TAMPERE UNIVERSITY OF TECHNOLOGY

MARIA LEHTIVAARA

Analysis of different hepatotoxicity types by transcript profiling

Master of Science Thesis

Examiner: Professor Olli Yli-Harja
Examiner and topic approved in
the Faculty of Science and Environmental Engineering Council
meeting on 15 August 2012

TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO
Biotekniikan koulutusohjelma
Maria Lehtivaara: Erilaisten hepatotoksisuustyyppien analyysi transkriptioprofiloinnilla
Diplomityö, 72 sivua, 22 liitesivua
Syyskuu 2012
Pääaine: Laskennallinen systeemibiologia
Tarkastajat: Olli Yli-Harja
Avainsanat: Hepatotoksisuus, varhaisen vaiheen lääkekehitys, transkriptioprofiili, mikrosirun data-analyysi

Tutkimusprojektin tavoitteena oli selvittää, onko geenien ilmentymisen muutoksista mahdollista tunnistaa erityyppisiä maksan kudospoikkeamia välittämättä niitä aiheuttaneista toksiineista. Toisena tavoitteena oli tutkia, voitaisiinko geenimuutosdataa käyttää tulevaisuudessa referenssinä ennustettaessa mahdollisia kudospoikkeamia potentiaalisilla lääkeainemolekyyleillä käsitellyissä näytteissä. Tätä tarkoitusta varten käytettiin japanilaisen toksikogenomiikkaprojektin tietokannan (TG-GATEs, julkaistu 25.2.2011) DNA-sirudataa rotan maksasta *in vivo* mitatuista geeni-ilmentymistä. Mahdollisuutta käyttää pienempiä annoksia ja lyhyempiä koeaikoja resurssien säästämiseksi tutkittiin vertailemalla eri altistumisaikoja ja annoskokoja.

Eroja eri kudospoikkeamaryhmien välillä tutkittiin pääkomponenttianalysilla (PCA) ja KNN ristiintestausmenetelmällä. Kunkin ryhmän normaalista poikkeavasti ilmentyneet geenit ja näissä poikkeuksellisen usein esiintyneet geeniontologiat listattiin, ja geenilistoja vertailtiin ja tutkittiin tarkemmin toksisuuden syiden ja mekanismien selvittämiseksi. Aika- ja annosvastetta analysoitiin samalla menetelmällä. Datan käyttöä kudosmuutosten ennustamisessa tutkittiin suorittamalla joitakin vertailuja sekä tunnettujen että mahdollisten lääkeainekandidaattimolekyylien avulla.

Tutkimuksen teoriaosuudessa käsitellään DNA-sirujen tekniikkaa ja datan käsittelyä sekä esitellään toksikogenomiikkaa alana sekä TG-GATEs -tietokantaa. Analyyseissa on käytetty R-ohjelmointikieltä, Bioconductor-paketteja sekä Ingenuity IPA -ohjelmistoa. Räätälöityjä CDF-tiedostoja analyysissa käytettiin mahdollisimman ajanmukaisten ja tarkkojen tulosten saamiseksi.

Tulokset osoittavat, että toiset kudosmuutosryhmät erottuvat kontrollinäytteistä paremmin kuin toiset. Näytteiden hajontaa selittävät paitsi analyysin muuttujat (löydösten vakavuus, altistusaika ja annoskoko) myös erilaiset muutoksia aiheuttavat mekanismit. Tulosten perusteella näyttää siltä, että geeni-ilmentymän muutoksia voitaisiin käyttää yhtenä menetelmänä lääkeaineiden toksisuutta ja sen mekanismeja arvioitaessa. Annos- ja aikavastetta tutkittaessa huomattiin, että näytteitä tarvitaan tulevaisuudessakin eri annostuksella ja eri altistusajalla tehdyistä kokeista, sillä pelkät yksittäiset datapisteet saattavat geeni-ilmentymistä tutkittaessa olla harhaanjohtavia.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Biotechnology
Maria Lehtivaara: Analysis of different hepatotoxicity types by transcript profiling
Master of Science Thesis, 72 pages, 22 Appendix pages
September 2012
Major: Computational Systems Biology
Examiner: Olli Yli-Harja
Keywords: Hepatotoxicity, early-stage drug development, transcript profiling, microarray data analysis

The aim of this project was to study whether it is possible to identify histopathological changes in liver based on gene expression profiles, regardless of the toxic causing the change. Another goal of the project was to study whether gene expression data can be used as a reference when predicting possible histopathologies in future samples treated with possible drug candidates. For this purpose, data from *in vivo* microarray gene expression studies of rat liver from Japanese Toxicogenomics Project database (TG-GATEs, published in February 25, 2011) was used. Possibility of using smaller doses and shorter experiment times in order to save resources was studied by comparing different exposure times and dose levels.

Separation between different histopathological groups was studied using Principal Component Analysis and K-nearest neighbor cross-validation. The differentially expressed genes and enriched gene ontologies from each group were listed, and the gene lists were studied further and compared to observe possible causes and mechanisms of toxicity. Time and dose responses were analyzed similarly. Some comparisons using well-known and possible candidate drug molecules were made in order to test the predictive use of the data.

In the theory section microarray technology and data analysis are explained, and the field of toxicogenomics and TG-GATEs -database are presented based on articles from scientific journals and books. Analyses are performed using R-language, Bioconductor-packages and Ingenuity IPA -software. Customized CDF-files were used in analysis to provide most current and accurate results.

It was noticed that while other histopathologies are clearly separated from the controls, others are more spread out. Spreading is likely to be caused by number of different mechanisms causing the toxicity in some groups, but also due to different analysis factors (severities of the findings, exposure times and dose levels). It seems that the data and gene expression profiles can be used as references when estimating possible toxicity of a drug molecule and the mechanisms of it. Some dose and time response was noticed. Due to the complex nature of gene expression, samples treated with different doses and exposure times are needed to see the direction of the development: single values from a single data point can be misleading.

PREFACE

The project reported in this Master's thesis was prepared during the spring and summer 2012 at Orion Oyj in Espoo in the Research and Development sector, in Biomarkers Group. The group is headed by Elina Serkkola. The main instructor and supervisor during the project was Daniel Nicorici.

I wish to express my gratitude to Elina for giving me a valuable opportunity to get acquainted with the drug development process and to see related research fields in Orion. I thank also Daniel for being so kind and helpful with the instructions and making the project so easy for me to approach. I am also thankful to Juha Kaivola and Kristiina Haasio, who helped me several times with the most varied challanges I came up with. Elina, Daniel, Juha and Kristiina were never too busy to guide me or discuss with me about things more or less related to the topic. I got valuable advices also from Ken Lindstedt and Sami Virtanen. Special thanks belong to Anneli Mustonen, who patiently explained me the rules and practices in the company.

I also thank Olli Yli-Harja who helped me with all the practical issues and examined this thesis. Thanks to Tommi Aho, Matti Nykter and Jari Yli-Hietanen for borrowing me books, computer and space, and otherwise supporting me with the whole project.

Thanks to my family, friends and co-workers for all the support. Thanks also to my very bestest brother Jaakko for proofreading this thesis.

Espoo, September 2012

Maria Lehtivaara

TABLE OF CONTENTS

ABBREVIATIONS AND NOTATIONS

ALT	Alanine aminotransferase, often measured from blood samples to get information of events taking place in liver.
ALP	Alkaline phosphatase (see ALT).
AST	Aspartate aminotransferase (see ALT).
BP	Biological Process (term used in GO).
CC	Cellular Component (term used in GO).
CDF	Chip Description File (CDF) contains information about the layout of the microarray chip used. It describes the link between probes and probesets, and identifies PM and MM probes and control probes.
cDNA	Complementary DNA, one-stranded DNA molecule synthesized from mRNA with reverse transciption. Does not contain introns.
CEL	CEL files contain measured intensities and locations for microarrays. Generated by the GCOS software.
CV	Coefficient of variation: the ratio of the standard deviation to the mean.
DAT	DAT file contains the scanned microarray image. Generated by the GCOS software.
DNA	Deoxyribonucleic acid, molecule containing genetic information.
EPA	Environmental Protection Agency of United States.
FC	Fold-change.
FDA	Food and Drug Administration, agency in United States supervising i.e. pharmaceutical drugs.

FDR	False discovery rate control, a statistical method used in multiple hypothesis testing to correct for multiple comparisons.
GCOS	GeneChip Operating Software by Affymetrix: used to scan and save the image of the microarray (DAT-file) and to compute the intensity values (CEL-file).
GGA	Groung glass appearance, a term used in histopathology.
GO	Gene Ontology: describes in which biological process, cellular component or molecular function the gene is present.
HC	Hiearchical clustering algorithm.
IPA	Ingenuity Pathway Analysis.
KNN	K-Nearest Neighbor classification algorithm.
LDH	Lactate dehydrogenase (see ALT).
MAQC	MicroArray Quality Control Consortium evaluates microarrays, methods of using them and the data analysis process.
MAS5.0	A microarray data analysis software by Affymetrix.
MF	Molecular Function (term used in GO).
MM	Miss-match probe, used in Affymetrix microarray chips to detect unspecific binding. The sequence of the probe is same as in PM probe, but one base in the middle of the probe is changed.
mRNA	Messenger-RNA, ribonucleic acid transcriped from DNA in nucleus and translated into protein sequence in ribosomes.).
PCA	Principal Component Analysis.
PCR	Polymerase Chain Reaction, a biochemical technology used to amplify small amounts of DNA.

PM	Perfect match probe, used in Affymetrix microarray chips to detect binding of cDNAs from a test sample.
R	R -programming language.
RMA	Robust Multiarray Analysis.
SNP	Single-nucleotide polymorphism, DNA variation of single nucleotide across the members of species.
TBIL	Total bilirubin (see ALT).
TG-GATEs	The gene expression database produced during the TGP project of Japanese government and private pharmaceutical companies. Short for TG-Genomics-Assisted Toxicity Evaluation system.
TGP	Japanese Toxicogenomics Project, performed by National Institute of Health Sciences, 15 pharmaceutical companies and the National Institute Biomedical Innovation (NIBIO) in Japan.
UniGene	UniGene is a internet tool that computationally identifies transcripts from the same locus, analyzes expression by tissue, age, and health status, and reports related proteins (protEST) and clone resources.

1. INTRODUCTION

In order to eliminate the potentially harmful chemicals at the early stage of the drug development process, it would be extremely useful to be able to identify the molecular changes caused by a toxic which is known to cause pathological changes in certain organs. These molecular changes can be detected in gene expression profiles, protein synthesis and metabolism of an organism. [69, 35, 67, 56]

Gene expression profiling can be defined as measuring of the activity of genes in certain cell(s). These measurements are usually done with microarrays, small chips containing probes for each gene of interest. Gene expression activity reflects the situation in the cells(s), since it is regulated based on the signals the cell receives. By regulating gene expression the cell adjusts to the changed situation by producing different amounts of proteins that function for example as enzymes, signal molecules or building blocks of larger complexes. Gene expression profiles can be considered as the most sensitive measure, and therefore gene expression patterns caused by toxic chemicals are extensively studied. Data generated by these toxicogenomics studies can be found in several large, publicly available databases. [69, 35, 67, 56]

Toxicity is often studied by comparing samples treated with toxic compounds to controls and hence predicting possible biomarkers of toxicity [68]. However, the comparison is then done without actual knowledge about the phenotype of the samples, and this might cause unwanted variation in the results. Thus, the comparison between known toxic conditions (that is, the effects of the toxin are seen on the phenotype for example as histopathological findings) and controls should provide more reliable foundation for predictions.

The aim of this project is to study whether it is possible to identify histopathological changes in liver based on gene expression profiles, regardless of the toxicant causing the change. For this purpose, data from *in vivo* gene expression studies of rat liver from TG-GATEs database [47] is used. The database contains microarray test results done with 150 or so different chemicals (most of which are medical drugs) and corresponding control studies. Several dosages and exposure times were studied. Affymetrix GeneChips were used as microarrays. The livers of the test animals were also analyzed. Based on these histopathological findings, the microarray results can be grouped, and the possible congruence of the most severe and common pathological changes can be studied. [67, 69, 68]

Histopathological groups are studied by detecting the differentially expressed genes between each histopathological group and the controls. These genes can then be analyzed further based on their ontologies. Principal component analysis and k-nearest neighbor classification can be used to highlight the differences between the histopathological groups and which histopathological groups are distinguishable from each other based on the gene

expression data. Tools from Bioconductor-project [10] are used in these analyses, and analysis codes are written in R-language [16]. Some of the gene lists generated using Bioconductor and R were uploaded and analysed further using Ingenuity pathway analysis (IPA, Ingenuity Systems, Redwood City, CA) [64]. Due to the different grouping of the microarray experiments, batch effect correction is needed before analysis. Customized CDF-files (provided by Microarray Laboratory of Molecular and Behavioral Neuroscience Institute of University of Michigan) [45] are used in analysis to provide most current and accurate results.

First, the theory behind microarrays and their analysis, and the field of toxicogenomics are presented. Current situation in the field is briefly outlined. Materials and methods are described in detail. Results of the analyses are reported, and their possible use and relevance is discussed together with other issues related to toxicogenomics and drug development, as well as possible ideas for future studies.

2. THEORETICAL BACKGROUND

In this Chapter the field of toxicogenomics and its use in drug development are described, as well as the current situation in the field. Microarray technologies as well as data analysis methods and tools are presented in detail. Also the histopathological terms used in this project are explained. Focus is in TG-GATEs database and Affymetrix arrays due to their central role in this project.

2.1 Toxicogenomics in Drug Development

The amount of new drugs that enter the pharmaceutical market is decreasing, even though the technology is advancing: one estimation of the success rate of a new drug is 1 in 10 000. Some state that the reasons for this are more tight regulations and the difficulty of clinical trials, but one important reason might be the inconsistency between preclinical and clinical tests and results. Many drug candidates are dropped out due to their toxicity, and thus effective ways for testing possible toxicity are valuable to pharmaceutical research. [69]

Toxicogenomics can be determined as a study of changes in gene expression caused by a toxic substance. In modern drug development it is important to eliminate potentially harmful drug candidates as early and cheaply as possible [39]. Modern high-throughput techniques, such as microarrays, may offer an opportunity to do this effectively [39]. Here, the toxicogenomics and ongoing projects in this field are presented. Special attention is given to Japanese Toxicogenomics Project (TGP) and its database (TG-GATEs) which is used in this study.

2.1.1 Toxicogenomics

In the past it was common for drug development to contain considerable amounts of test, where different chemicals were used on animals carrying artificially caused diseases. Besides ethical, time and money related problems, this method suffers from differences between test animals and human targets. After the sequencing of genomes of many common test animals and humans it has been possible to target the drug research on well-known molecular processes. Despite of this, many drugs advance to later stages of development before their toxicity is noticed. [67]

In order to eliminate the potentially harmful chemicals at the early stage of the drug development process, it would be extremely useful to be able to identify the molecular changes caused by a toxic known to cause pathological changes in certain organs. These molecular changes can be detected in gene expression profiles, protein synthesis and metabolism of

the organism. The first mentioned can be considered as the most sensitive measure, and therefore gene expression patterns caused by toxic chemicals are a widely studied field. The identification of these predictive biomarkers for toxicity during the pre-clinical stages of drug development is of great importance to pharmaceutical companies. [67]

Toxicogenomics is determined as the application of microarray technologies to toxicology studies, and it is widely used in pharmaceutical studies to identify chemicals with potential safety problems [35, 67]. Besides identifying biomarkers of toxicity in order to predict the hazardous chemicals, it is also used to study the molecular mechanisms of toxicity [35, 67]. It provides understanding of molecular changes caused by a disease or a treatment [35], and can be used to detect relationships between changes in gene expression and end-point data (for example histopathological findings, clinical chemistry) [67]. One ultimate goal of toxicogenomics is to aid in risk assessment together with more traditional methods used in drug development [67].

Toxicogenomics is considerably cheap, easy and fast. It produces lots of data, and lots of references and databases exist and are becoming available. Microarray methods and the ways of reporting the data are being standardized. Food and Drug Administration of the United States (FDA) has identified toxicogenomics as a "key opportunity in advancing personalized medicine" and together with Environmental Protection Agency (EPA) they have presented guidance documents to encourage scientific research for using toxicogenomics data in drug development, medical diagnostics and risk assessment. [56]

2.1.2 Ongoing projects

Due to vast amount of data generated by toxicogenomics studies, large and well-designed databases are needed. The data is generated and presented in standardized format (MIAME, see Chapter 2.2.2), which allows for its effective use. [67]

There are several public databases, such as Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo), ArrayExpress (www.ebi.ac.uk/microarray-as/ae/), Center for Information Biology Gene Expression (CIBEX; http://cibex.nig.ac.jp/), EDGE (http://edge.oncology.wisc.edu/edge3.php), Chemical Effects in Biological Systems (CEBS; http://cebs.niehs.nih.gov/cebs-browser/), dbZach (http://dbzach.fst.msu.edu) and Comparative Toxicogenomics Database (CTD; Laboratory; http://ctd.mdibl.org).

In addition there are few collaboration projects such as European InnoMed PredTox (http://www.genedata.com/lp/innomed-predtox.html), American Liver Toxicity Biomarker Study and Japanese Toxicogenomics Project (TGP; http://wwwtgp.nibio. go.jp/index-e.html), which all are collaborations of pharmaceutical companies and academic institutions. [67]

2.1.2.1 Japanese Toxicogenomics Project

The TGP was performed in 2002 – 2007 by National Institute of Health Sciences, 17 pharmaceutical companies and the National Institute Biomedical Innovation (NIBIO) in Japan. The original name of the project was *Construction of a forecasting system for drug*

safety based on the toxicogenomics technique and related basic studies. The actual work was performed in NIBIO. The primary goal of the project was to create an extensive gene expression database called TG- GATEs (Genomics-Assisted Toxicity Evaluation system) using Affymetrix GeneChips and 150 different chemicals, most of which are medical drugs. List of these compounds is presented in table 3.1 on page 25. [67, 69]

Rat, which is often used as model organism in toxicological studies, was used as a test animal. The gene expression was measured from liver samples of the test animals, since most toxic compounds affect liver cells (hepatotoxicity). Also, liver has relatively homogeneous composition of different cell types, and homogeneous samples mean less unwanted variation in the gene expression measurements. Another organ used was the kidney. Both *in vivo* and *in vitro* studies were performed. In this Chapter we focus on the liver *in vivo* studies, as the data used in this project is from these studies. The aim of the project is to perform same *in vitro* tests to human hepatocyte cells as well, so that interspecies bridging could be considered. [67, 69]

In *in vivo* studies, the test animals were treated either with single dose or multiple doses. In the single-dose study, testing was done in many different time points and with different dose levels. In repeated-dose study there were many different treatment periods and different dose levels. Body weight, general symptoms, hematology, blood biochemistry and organ weight were collected from each test animal. Besides the gene expression analysis, histopathological examination was performed to liver and kidney samples. The protocol used is described in more detail in Chapter 3.2. [67, 69]

At the moment, the TG-GATEs database is published, and it is publicly available at http://toxico.nibio.go.jp/. Some of the data intended is still missing from the database. The TGP has reached its second stage, TGP2. The goals of this continued project are to find genomic biomarkers for toxicity prediction, to discover bridging of differences between species (essentially between rat and human) and apply toxicogenomics and genomic biomarkers as regulatory part in drug safety assessment. [67, 69]

2.2 Microarrays

Microarrays are a high-throughput biochemical technique used widely in gene expression analyses. They were developed in early 1990s, and their scope and the techniques related to them are continuously developed further. Currently, several thousand scientific articles related to or using microarrays are published each year [15]. The technology, different types of arrays and the data analysis process are briefly explained here. A closer look is taken on the Affymetrix arrays, which were used when creating the TG-GATEs database.

2.2.1 The Central Dogma of Biochemistry

A phenotype of an organism is determined by its genes. Each cell in an organism contains the same set of genes in chromosomes that consist of deoxyribonucleic acid (DNA). (Some exceptions exist: for example red blood cells do not contain DNA.) DNA is a long molecule consisting of two complementary strands. These strands consist of several monomers which

contain a pentose sugar part, a phosphate part and a nitrogenous base part, which can be either adenine (A), thymine (T) cytosine (C) or guanine (G). These molecules in the two strands pair up selectively to A-T and C-G pairs (Watson-Crick base pairing). The order of these basepairs is called the gene sequence, and this sequence is responsible for the behavior of a gene. DNA contains informative parts (exons) and parts that are not used (introns). [46, 62, 73]

When a gene is expressed in a cell, it is first transcribed to a mRNA (messenger ribonucleic acid). The mRNA can be considered as a one-stranded version of DNA, with one additional hydroxy (-OH) group in the pentose ring and thymine replaced by similar base called uracil (U). Thus, the sequence of the DNA molecule is coded to mRNA molecule, which is then transported out from the nucleus of the cell. The mRNA is spliced so that only exon-parts of DNA-code remain. [46, 62, 73]

After this the mRNA can be translated to aminoacid sequence of a protein. Translation takes place in ribosomes, and it is enabled by transfer RNAs (tRNA). One tRNA can bind to three bases (a *codon*), and it carries one amino acid complementary to those three bases. For example, codon CAG codes for glutamine (one of the 20 amino acids), ATG is the "start" codon and TAA, TAG and TGA are "stop" codons. These codons are universal: the same code applies in all species. [46, 62, 73]

The process of transcription and translation form the basis of the central dogma of biochemistry (see Fig. 2.1). Proteins are then responsible for the behavior of cells as they function for example as enzymes, signal molecules or building blocks of the cell. [46, 62, 73]

One must keep in mind that the situation is often much more complex than the straight forward DNA-mRNA-protein-function -scheme described above and that there are many exceptions. Instead, genes and proteins often form complex networks and affect each others behavior in many ways. Also the dynamics of transcription and translation as well as the lifetimes of these compounds can vary. Therefore, a small change in the expression of a certain gene can have a great influence on the cell, whereas an even larger change in another gene's expression might have very little or no effect on the cell. [46, 62, 73]

During transcription, *alternative splicing* takes place: from a single DNA strand, several different mRNAs can be produced be choosing different exons. Some of the mRNA molecules are never translated to proteins, and some mRNAs code for proteins that are parts of larger protein complexes, so that the single protein is not functional on its own. Many proteins also go through *post-translational modifications* before they are fully functional. [46, 62, 73]

The expression process described above is often triggered by some signal coming to the cell. These signals can be mediated via signal molecules such as hormones, and they reflect the processes taking place in the environment of the cell or elsewhere in the organism. Thus, the expression patterns (which represent the amount of different mRNA molecules produced in the cell) reflect the situation of the cell itself, the organ of which the cell is a part of, and the whole organism. [46, 62, 73]

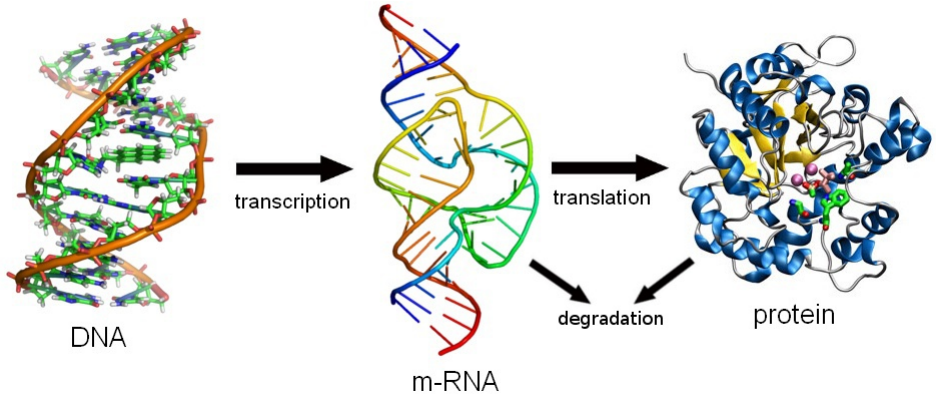


Figure 2.1: Schematic illustration of the central dogma of biology.

2.2.2 Microarray Technology

The mRNA molecules can be isolated from a cell sample with biochemical methods. Since mRNAs degrade easily, they need to be reversibly transcribed into single stranded DNA and then amplified with automated polymerase chain reaction (PCR). Since the DNA strands bind to each other specifically via the basepairing, appearance of a particular DNA -sequence in a sample can be tested by allowing it to pair up with a short piece of DNA with the complementary sequence (called as probe). By binding the probe to a surface and labeling the DNAs in the sample with a dye, the binding can be discovered visually. Microarray technology takes advantage of these basic biochemical methods. [46]

Microarrays are glass, silicon or plastic chips of few cm^2 that can contain from hundred to many thousands of test sites [25]. A test site is an area of 10-500 microns that contain probes, which can be oligonucleotides (<50 bases long nucleotide chains) or larger DNA or RNA fragments [25]. Probes are used as targets into which the reporter molecules are hybridized. These probes are attached to the chip covalently or noncovalently by either synthesizing them *in situ* or immobilizing them to the test sites [25].

After the hybridization and washing of the chip the reporter molecules are detected using a signal generated by the binding of the molecule to the target. This signal can be for example fluorescent, chemiluminescent, colorimetric or a radioisotope. The chip is scanned or imaged to save the signal patterns, and the resulting data is then analyzed [25].

When the slide is scanned, the scanner excites it with a laser and detects the resulting photon emissions from the fluorescently labeled molecules hybridized to the probes in the slide. The detected photon emissions are converted to 16-bit intensity values. There are several scanners available commercially. [54]

To avoid confusion and enable comparison between research work done in different groups, international standards have been developed by Microarray Gene Expression Data (MGED) Society (http://www.mged.org). They have published the Minimum Information About a Microarray Experiment (MIAME), which describes what kind of information should at least be submitted with microarray results, and how should this information be presented. [37]

2.2.3 Different Types of Microarrays

There are the two main types of microarrays: robotically spotted and *in situ* photolithographically hybridized arrays. The first ones are sometimes referred as cDNA microarrays because the nucleic acids immobilized to the chips used to be PCR amplified cDNAs from libraries. The latter ones can be called as oligonucleotide arrays (or oligoarrays) because they use several shorter oligonucleotides to represent a certain gene. These oligonucleotide arrays are sometimes referred also as Affymetrix arrays after the well-known commercial manufacturer. Nowadays oligonucleotide arrays are also prepared with robotically, so these terms can be slightly misleading. [37]

Spotted arrays use so-called comparative hybridization, which means that two samples (i.e. control vs. experiment, healthy vs. diseased) are labeled with two different dyes (usually red and green fluorescent dye) and then hybridized to the same array. Therefore the color of the spot implies whether another sample had higher amount of the mRNA in question (the spot is either red or green), or whether the expression levels were somewhat equal in both samples (yellow spots). [37]

Oligonucleotide arrays use only one color for labels, and each sample is hybridized to a separate array. Then the intensity of the spot reflects the level of gene expression, and the comparison between the samples is made afterwards. In Fig. 2.2 the difference between the use of spotted arrays and oligonucleotide arrays is presented. [37]

The differences in the array design are reflected to their analyses as well. In spotted arrays the test sites are (as the name says) round in shape, whereas in oligoarrays they are rectangular. This causes different concerns to background adjustment during the analysis. With spotted arrays the possibly different behavior of the labels needs to be taken into account during the analysis. With oligonucleotide arrays the normalization between different arrays might be troublesome. Both techniques are troubled with unpredictable differences in results due to the different handling of arrays, samples and dyes. Since the TG-GATEs database used in this project is build using Affymetrix oligonucleotide arrays, next chapter focuses on them. [26, 50]

2.2.3.1 Affymetrix GeneChip arrays

There are several companies producing microarrays and related systems and applications. Affymetrix Inc. is one of them. Affymetrix 1.6 cm^2 Genechips are used by thousands of researchers. They are manufactured with *in situ* photolithographic process, and they differ from red/green channel spotted arrays as they have only one channel and that the test sites are rectangular in shape. Due to use of only one channel, only one sample can be put on each slide; separate control slides are needed. Affymetrix mass-produces arrays for several organisms: human, mouse, rat, *arabidopsis*, *Drosophila*, yeast, zebrafish, canine and *E. coli*. [29]

In each test site of 24 x 24 μm there is a set of 11-20 pairs of probes with same sequence. There are $10^6 - 10^7$ copies of each probe. Each probe is (14-)25 bases in length. The probe pairs contain a perfect match (PM) and a mismatch (MM) probe. MM probes have same

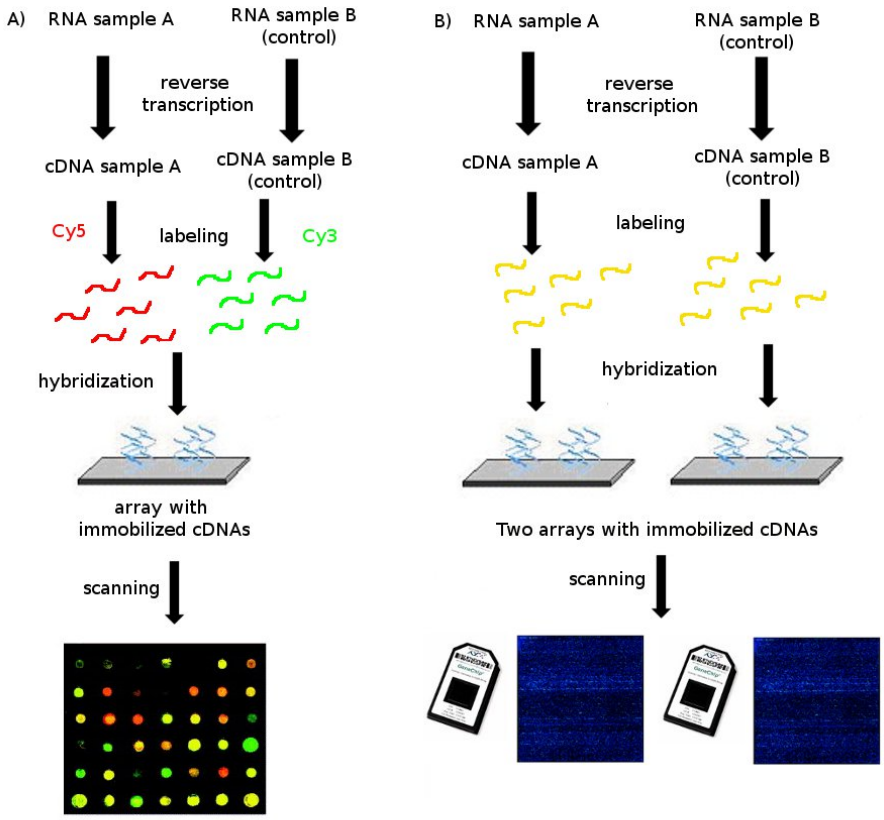


Figure 2.2: Illustration showing the principles of two-colored spotted arrays (A) and single color oligonucleotide arrays (B), such as Affymetrix GeneChips.

sequence as the PM probes, but the 13th (the middle) base is changed to its Watson-Crick complement in order to measure non-specific binding (see Figure 2.3). [29, 54, 3]

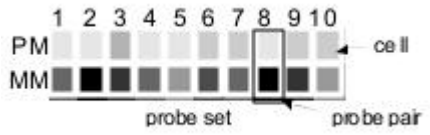


Figure 2.3: A schematic illustration of probesets and MM and PM probes [3].

For analysis of the scanned microarray data, Affymetrix has its own software tool: GCOS (GeneChip Operating Software). This software saves the scanned image and computes the intensity values. The scanned image is saved as `.dat` -file and the intensity data computed from this image is saved as `.cel` -file. For each probe set, a Signal value and a Detection call are computed. The Detection call classifies the samples as present, absent or marginal based on p-values. Signal value is calculated using the One-Step Tukeys Biweight Estimate. These methods are described in more detail in Chapter 2.3.3.2 and in [1].

The Affymetrix GeneChips also contain several probes for quality control: poly-A-controls (genes *dap, lys, phe, thr,* and *trp*) can be used to monitor the entire target labeling process and the genes *BioB, bioC* and *bioD* for monitoring the hybridization. For these controls, there are Control Kits available, that can be used to create samples with spiked concentrations of these genes. To assure RNA sample and assay quality, β-*actin* and *GAPDH* are used as internal controls. Their signal values from the 3' probe sets and 5' probe sets are compared: this value reflects the degradation of RNA. [1]

The Affymetrix GeneChips were designed by Steve Fodor and colleagues in the early 1990s with the methods used by computer microchip manufacturers. To prepare an Affymetrix GeneChip, the chip is first modified with silane reagent to provide hydroxyalkyl groups as initial synthesis sites, which are then extended with shielded linker groups. With masks and exposition of the chip to light, those protecting groups can be selectively removed. These un-protected groups can then be coupled with protected nucleoside phosphoramidite monomers, which can then again be coupled to the substrates using phosphoramidite DNA synthesis. These cycles of removing of the shielding photolabile molecules and addition of nucleotides can be repeated, and thus any nucleotide sequences can be build on the array. [25]

2.2.4 Applications

There are many applications for DNA microarrays. Usually these contain either gene expression analysis or screening for single nucleotide polymorphisms. Therefore microarrays are used in basic molecular biology analyses and in genetic research, but also in fields such as pharmacogenomics research, infectious and genetic disease and cancer diagnostics and in forensic and genetic identification purposes. Microarray technologies are likely to be developed further and gain importance also in other fields due to their economical and

Figure 2.4: A photo showing the size of a Affymetrix GeneChip. [17]

high-throughput benefits. [25]

2.3 Microarray Data Analysis

Microarrays are a high-throughput technique: they generate huge amounts of data in a short amount of time. Hence effective methods are needed for the data analysis, and several different methods have been developed lately for each step of the analysis. The methods used can also have a great influence on the results. Here the basic steps of the data analysis are described, focusing on (single color) oligonucleotide arrays used in this project. Bioconductor -project is also introduced, and some basic statistics. [50]

2.3.1 Statistics

In this Chapter some basic statistics concepts are presented to cover the methods used in the data analysis. In data analysis tasks one often compares two groups to see whether they differ. This brings one to the essential question in statistics: what is "significantly different"?

To compare two hypotheses, a **test statistic** and the corresponding **p-value** are computed. There are hundreds of statistical tests that can be used. Choice of the statistical test depends on the number of groups to be compared: for two groups, **t-test** and Mann-Whitney (U) -test are used most often. For more than two groups **analysis of variance (ANOVA)** or Kruskal-Wallis test are often used. [27]

Hypothesis pair consists of null hypothesis H_0 and alternative hypothesis H_1: usually H_0 is decided to mean that there is no difference between the groups compared, and H_1 that they do differ. The p-value can be determined as the risk that we reject H_0 when we should not. Thus some relatively small cut-off for p-value is needed: usually 0.05 or 0.001 are used. Larger data can often cause smaller p-values. [27]

T-test compares the means of the two groups (Welsh's t-test). The test statistic T is

Table 2.1: Example of a contingency table.

	Differentially expressed genes	Normally expressed genes	Total
Belongs to certain pathway	a	b	a + b
Does not belong to that pathway	c	d	c + d
Total	a + c	b + d	a + b + c + d = n

computed as follows:

$$T = \frac{X_i - Xj}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}} \tag{2.1}$$

where X_i and X_j are the means of the two groups, s_i^2 and s_i^2 are the variances of the groups and n_j and n_j are the number of samples in the groups. This test does not assume that the variances of the groups are equal: if this assumption is made, student's t-test can be used. Student's t-test is similar to Eq. 2.1, but the denominator is multiplied with s:

$$s = \sqrt{\frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}} \tag{2.2}$$

ANOVA tests for significant difference between means of several groups: if we are comparing two groups with one-way ANOVA, the results will be same as with the t-test. [27]

As the p-value presents the probability of falsely rejecting the null hypothesis, some methods are developed to estimate the p-value cut-offs, i.e. what is "sufficiently improbable". Perhaps the most used method for this is the **false discovery rate** (FDR). FDR can be explained as the expected proportion of false positives among the differentially expressed genes. One of the most used methods for this is the Benjamini and Hochberg -correction. [5, 24]

Fisher's exact test is used to compute the probability of getting the observed data. For this purpose, hypergeometric distribution is used. For a 2x2 contingency table (Table 2.1) the probability would be:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \tag{2.3}$$

2.3.2 Bioconductor Project

Bioconductor (http://www.bioconductor.org) is an open source project for computational biology. The project began in 2001, and the first packages were released in 2002. The main focus of the project is to deliver high-quality infrastructure and tools for expression analysis in R. Bioconductor is an open development initiative, where the users are encouraged to become developers. Some of the packages and functions that are used in this project are

described here. [14]

2.3.3 Basic Steps

The data analysis process can be divided into three basic steps: image analysis, preprocessing and normalization, and data harvesting. For each step, there are several possible methods. Image analysis is often done with a special software provided by the manufacturer of the microarray, while rest of the steps are usually performed by the researcher. [19, 27]

There are several opinions on which order the analysis steps should be taken [27]. In Fig. 2.5 one possible data analysis pipeline is shown. In this project, this kind of work flow was followed.

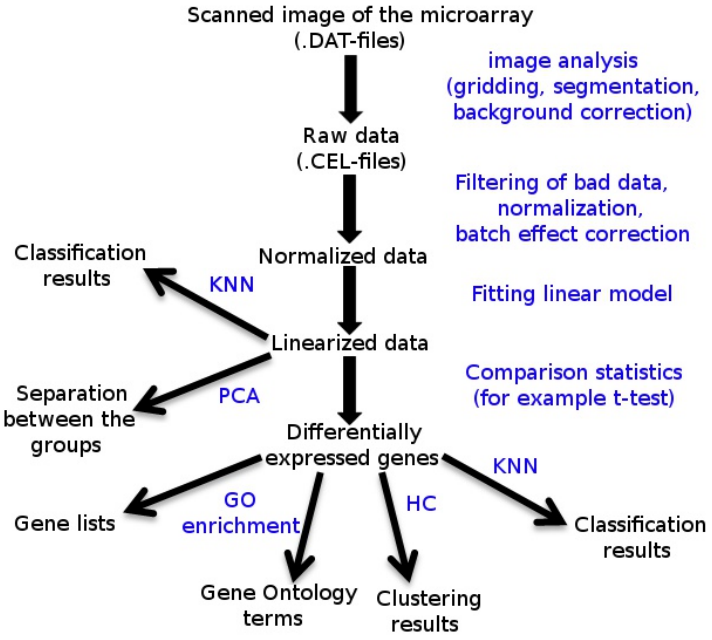


Figure 2.5: Flow chart of a microarray data analysis process.

2.3.3.1 Image Analysis

The very first step of data analysis after scanning of the slide is the image analysis. First, a grid is fitted over the image of the slide in order to separate the test sites from each other. Some slides (i.e. Affymetrix slides) contain alignment marks to make this easier.

After gridding, segmentation is performed to differentiate the background pixels from the actual foreground pixels. There are different methods for this, as well as for the background correction. Usually coefficients of variation (CV) are used to determine the optimal outlines for the grids. For background correction the slide is divided to smaller squares (16 by default in the Affymetrix software), and the lowest 2 % of the intensity values in this area are averaged and used as a background value on that area. [25, 54, 3]

2.3.3.2 Preprocessing and Normalization

The choice of **preprocessing** method can have important effect on the results, especially with oligonucleotide expression arrays such as Affymetrix GeneChips. Preprocessing can be divided into three steps: background adjustment, normalization and summarization. There are several functions available to perform each of these steps separately, and functions combining all of them. [19]

There are three basic schemes for **background correction**: ideal mismatch -method, MAS 5.0 background correction and RMA convolution. MAS 5.0 is an software by Affymetrix [28, 3]. The background adjustment method used by MAS 5.0 is described above in 2.3.3.1. Originally MAS 5.0 used also (log) PM-MM values as expression values, but this was noticed to generate noise to observations with low signal strengths. [19, 49]

Ideal mismatch method uses the MM probes as originally intended by the developers of the Affymetrix CeneChip: their intensity values are subtracted from the values of PM probes. However, since the MM value can often (about 30 % of the cases) be larger than PM value, a *specific background* (SB) value is calculated (using one-step biweight algorithm) to be used instead of MM value in these cases. Using SB we can determine the ideal mismatch (IM):

$$IM = \begin{cases} MM & \text{when MM} < \text{PM} \\ \frac{PM}{2^{SB}} & \text{when MM} \geq \text{PM and SB} > \tau_c \\ \frac{PM}{2^{\tau_c/(1+(\tau_c-SB)/\tau_s}} & \text{when MM} \geq \text{PM and SB} \leq \tau_c \end{cases} \tag{2.4}$$

where τ_c and τ_s are constants for tuning contrast and scaling (default values are 0.03 and 10, respectively). The background correction is done by subtracting IM -value from PM value (PM -IM). [19, 3]

RMA (Robust Multiarray Analysis) convolution, originally developed by Irizarry et al [30] as a part of the RMA method, does not use MM probes at all. Because of this, the variance of low abundance transcripts is reduced. Instead of using MM probes, the PM values are corrected using a model for the distribution of probe intensities across the array:

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}, \quad where \quad i = 1...I, j = 1...J, n = 1...n \tag{2.5}$$

where α_j a probe affinity effect, μ_i representing the log scale expression level for array i, and ε_{ij} representing an independent identically distributed error term with mean 0. Based on this model, an adjustment equation can be written:

$$E(S\|Y) = y) = a + b\frac{\phi(\frac{a}{b}) - \phi(\frac{y-a}{b})}{\Phi(\frac{a}{b}) + \Phi(\frac{y-a}{b}) - 1} \tag{2.6}$$

where $a = s - \mu - \sigma^2$ and $b = \sigma$, (μ = mean and σ^2 = variance). θ and Θ are standard normal density and distribution functions. S stands for the exponential signal component and Y for the observed intensity. [19, 49]

Normalization is the term used for manipulation of data in order to make it more comparable: microarray experiments usually contain multiple arrays, and it is of interest to remove the non-biological sources of variation. Again, there are several methods. [6, 19]

Scaling method, used by Affymetrix in their software, chooses one baseline array and scales all the other arrays so that they have the same mean intensity as the baseline array. Trimmed mean can also be used by removing the lowest and highest frequencies (usually 2% is removed from both ends). [19, 6]

Quantile normalization equalizes the probe intensity distributions between arrays in a set of arrays. The method is named after the idea of quantile-quantile plot, where a diagonal line can be seen if the two vectors plotted in it follow the same distribution. In n dimensions we would see a straight line in $\left(\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}}\right)$. The algorithm used to project the data points to the diagonal of the n dimensional quantile plot is the following:

1. Form matrix $\mathbf{X}$ with dimensions $p \times n$ (n is the number of arrays, p is the length of the array), where each array is a column.
2. Sort each column of $\mathbf{X}$ to get $\mathbf{X}_{sort}$
3. Compute the mean of each row and place this value to each element in the row to get $\mathbf{X}_{sort}^{'}$
4. To get $\mathbf{X}_{normalized}$, rearrange each column of $\mathbf{X}_{sort}^{'}$ according to the order in $\mathbf{X}$.

This method can be varied by estimating the distributions smoothly, but it can be noticed that the algorithm above performs rather well for high-density oligonucleotide data such as Affymetrix GeneChips. One problem in forcing the quantiles to be equal is that in the tails there might be probes having the same values in all arrays. In practice this is however unlikely, as the probesets expression is computed from multiple probe values. [19, 6]

Another normalizing methods are *cyclic loess*, *contrast normalization* and *non-linear methods*. The last-mentioned uses baseline array similar to scaling method, but the adjustments between arrays are non-linear [19, 6]. Cyclic loess normalizes arrays pairwise (with two-channel data, the two channels form the pairs) as described by Yang et al. [74]. The cycle of the pairwise combinations is continued until convergence, so this method is rather time-consuming. Contrast normalization method is faster and uses similar methods, described in detail in [4]. [6, 19]

Summarization is the last stage of pre-processing when working with Affymetrix data: it is the process of combining the multiple probe intensities for each probeset to get an expression value. Again, there are several methods available. Probably the most used way to summarize the probe intensities is the median polish algorithm, used for example in `RMA` -function.

Median polishing is an iterative procedure for extracting row and column effects in a two-way table using medians rather than means. In the algorithm, observations $y_{i,j}$ in a two-way table follow a model:

$$y_{i,j} = \mu + \alpha_i + \beta_j (+\epsilon_{i,j}) \tag{2.7}$$

where μ represents a grand effect, α_i and β_j denote the i:th row effect and the j:th column effect. ($\epsilon_{i,j}$ is the measurement error in the element (i, j).) In the least squares fit, these parameters are sought so that the row and column sums of the residuals are zero, but as the name says, median polishing method uses medians instead. The algorithm updates a table:

$$\begin{pmatrix} e_{1,1} & \cdots & a_{1,c} & a_1 \\ \vdots & \ddots & \vdots & \vdots \\ e_{r,1} & \cdots & e_{r,c} & a_r \\ b_1 & \cdots & b_r & m \end{pmatrix} \tag{2.8}$$

where initially $e_{ij} = yij$, $a_i = b_j = m = 0$. When updating the table, condition $y_{ij} = m + a_i + b_j + e_{ij}$ must hold. When updating the table, row and column sweeps are alternated: a row sweep means that for each row the median of columns 1,... ,c is subtracted from those columns and added to the last column, and vice versa for the columns sweep (=for each column, the median of rows 1,...,r is subtracted from those rows and it is added to the bottom row). These two steps are performed until the changes are considerably small (under a certain limit). [70]

2.3.3.3 Batch Effect

Microarray results can be affected by slight changes in non-biological variables. Such variables are for instance different types or lots of chips, reagents from different lots, different storage or shipment conditions of chips or reagents, washing conditions (temperature, ionic strengths), scanner calibration, different experimenters and changes in the environment (temperature, ozone levels) [11, 40]. Because of this, these variables are documented, and those tests that are done with same parameters (at one site over a short period of time using the same platform and experimenter) belong to same *batch*. [11]

The systematic error introduced when samples are processed in multiple batches is called *batch effect*. This error can be reduced by designing the experiments carefully, but the only way to absolutely eliminate it is to perform all the tests in a single batch. Understandably this is seldom possible with larger studies, and there are always some variables that are not under the control of the researcher. Therefore there are several programs available that can adjust the data according to the batches afterwards. [11]

However, according to [11], batch effect is rarely taken care of: in January-June 2010 only 10 % of the published microarray data papers addressed this issue. The batch effect can account as much as 50% of the observed variation in expression [11], so by adjusting

according to it the statistical power of the analysis can be increased and the underlying biological phenomena can be detected [38]. With batch effect correction it could in theory be possible to compare and combine microarray data sets even from different laboratories [38].

2.3.3.4 Identifying Differentially Expressed Genes

The main goal in many microarray studies is to define the differentially expressed genes. This means identification of those genes whose expression patterns differs according to their phenotype, treatment or any experimental condition. Here, those genes which are differentially expressed between control arrays and those arrays with histopathological findings (=phenotypical changes) are of interest. [19]

When discussing about differentially expressed genes, a question of determinition arises: what kind of value and cut-off should be used? An effort to standardize these issues was established by MicroArray Quality Control (MAQC) Consortium in 2006. MAQC is an effort managed by FDA (Food and Drug Administration of US) scientists and consisting of over hundred researchers at 51 different academic, government and commercial institutions. The goals of MAQC is to evaluate microarrays, experiments done on them and the data analysis process. A higher goal is to find the ways and limits of how microarrays can be used in decision making in regulatory affairs (such as drug selling licenses) and in the clinic. [15, 55, 56]

MAQC states in their publication [56] that often the statistical significance (represented by p-value) alone is used to determine the degree of differential expression, which has caused disagreement. Therefore they propose that also the actual measured quantity of differential expression (= fold change) should be used in identification. In their own studies, they use the following criteria: p-value < 0.001 and a mean difference greater than or equal to twofold. [56]

Use of pure fold change as a criteria is criticized, especially with a small dataset with only a few replicates, as it does not take into account the biological and experimental variation, which varies from gene to gene [19, 31]. This means that distributions of expression levels are not always normally distributed and that the distributions may vary from gene to gene [58]. Therefore there are several statistical tests available for more clever analysis. Some of these methods were compared by I. Jeffery et al. in [31], where they state that the *Empirical Bayes statistic*, the *Area Under the ROC curve* -method and *Rank product*s are the most accurate ways for identification of differentially expressed genes, and that these methods produce the most robust classifiers.

For linear modeling of the expression of each gene, Bioconductor has a package called `limma`. Usually the analysis of differential expression begins by fitting a linear model (function `lmFit` does this), which can be used to estimate the variability of the data and to distinguish the random variation. For this, a *design matrix* is needed. This matrix describes the difference of the RNA targets in each array, for example which arrays are controls and which are the test group. Hence each row in the design matrix corresponds to

an array and each column to a different source of RNA: in the example case, there would be two columns (controls and tests). *Contrast matrix* then again determines the wanted comparisons (=contrasts) between the coefficients: in the example case the comparison of interest would be as simple as disease-control. This simple comparison is equivalent to a one-way ANOVA (analysis of variance) for comparing two groups. [19, 58]

The linear model can be mathematically expressed as:

$$E\left[\mathbf{y}_j\right] = \mathbf{X}\alpha_j \tag{2.9}$$

where $\mathbf{y}_j$ is the expression data for gene j, $\mathbf{X}$ is the design matrix, α_j is the vector of coefficients. The contrasts of coefficients with biological interest are:

$$\beta_j = \mathbf{C}^T\alpha_j \tag{2.10}$$

where $\mathbf{C}$ is the contrast matrix. After `lmFit`, the next step is usually to use function `contrasts.fit`, which computes estimated coefficients and standard errors for a given set of contrasts. The idea is then to test the null hypothesis of $\beta_j = 0$. [19, 58]

The third step of an usual analysis is to use **eBayes** function to compute the moderated t-statistics, moderated F-statistic, and log-odds of differential expression. The function uses empirical Bayes shrinkage of the standard deviations of the probes towards a common value, which is done by including data for genes that are expressed at similar levels. [19, 58]

Thus, the moderated t-statistic is similar to a normal t-statistic, except that the standard errors have been shrunk towards a common value: it can be said that information is borrowed from the entity of genes. Therefore also the p-values computed from these t-statistics are a bit different, as they have more degrees of freedom. Log-odds present the probability of a gene to be differentially expressed, so that log-odd of zero means that there is a 50:50 change that the gene is differentially expressed. The moderated F-statistic combines t-statistic for all the contrasts into an overall test of significance of that gene. [19, 58]

The final step is to adjust the p-values. Function `topTable` contains several methods for the p-value correction. Most commonly used is the Benjamini and Hochberg -method [5], which is considered to give a good compromise between sensitivity and specificity. The method controls the false discovery rate (FDR, the expected proportion of false discoveries amongst the rejected hypotheses). [19, 58]

```
fit <- lmFit(Dataset, design)
contrast_matrix <- makeContrasts(disease-control, levels=design)
fit2 <- contrasts.fit(fit, contrasts_matrix)
fit3 <- eBayes(fit2)
results <- topTable(fit3, number = dim(Dataset)[1])
```

2.3.3.5 Gene Ontology enrichment

After determining the differentially expressed genes, one would usually like to know how they are related or connected to each other. This knowledge should make it easier to understand the biological phenomena taking place in the different experiments. One way to do this is by using Gene Enrichment Analysis and Hypergeometric tests.

Most eucaryotic organisms have same genes that control the biological core functions, and hence the biological knowledge can be transferred to other organisms as well. Thus it was inevitable to gather all the constantly changing and increasing knowledge about gene and protein roles in to a database: this was done by Gene Ontology (GO) Consortium. They determined three independent ontologies: biological process (BP), molecular function (MF) and cellular component (CC). [66, 48]

These terms form a tree of nodes: for each lower level term there is a higher level term. BP refers to a biological objective in which the gene is involved: a process often involves chemical or physical transformations, and contains one or more molecular functions. Examples of BP terms are ”cell growth and maintenance” (higher level) and ”translation” (lower level). Molecular function is defined as the biochemical activity of a gene product. Examples are ”enzyme”, ”ligand” (higher level) and ”adenylate cyclase” (lower level). Cellular component indicates the place in the cell where a gene product is active. Example terms are ”nuclear membrane” and ”Golgi apparatus”. [66, 48]

The question in gene enrichment is: do the differentially expressed genes belong to same gene ontology category? This issue can be solved with probabilities, i.e. by telling how likely it is that in a certain GO category there are this many interesting genes. Based on this, p-values are computed. Hypergeometric distribution is used to model the number of interesting genes in the GO category. This test is often called as Fisher's exact test.[24]

For this test, one has to determine the universe of genes. The choice of the universe has of course effect on the p-values: in [24] it is recommended to use all ”possibly interesting genes” as universe. This can mean the genes that were present on the slide, or the genes that have probes on the slide. The list of interesting genes is also needed, i.e. those genes that were differentially expressed (see section 2.3.3.4). Genes that are presented by more than one probe can cause problems in the analysis, and for the analysis to be functional, only one value per gene has to be chosen. [24]

With Bioconductor, Hypergeometric testing is performed with `hyperGTest` -function. The parameters are given to `hyperGTest` as a set called `GOHyperGParams`. The parameters are: gene universe, list of interesting genes, annotation data package used, which ontology to use (BP, CC, MF), p-value cut-off and the test direction (over or underpresentation of the GO terms). As a result, `hyperGTest` returns an object (called `HyperGResult`), which can be viewed as a summary, or which can be generated into a html -report including links to Gene Ontology -webpage [48, 8].

2.3.3.6 KNN

K-nearest neighbors (KNN) is a simple classifying algorithm, which classifies a sample based on the class of those k samples that are its nearest neighbors (if k=1, the classification is done based on the nearest neighbor). The neighbors with known class are called as training examples: they are usually vectors with one feature in each element and with a certain class label. For example, with gene expression data, each array can be considered as a sample and the probe values are its features. Ergo, a sample with unknown class is classified in the same class with those samples that have the most similar gene expression values (see Fig. 2.6). [61]

There are several methods for selecting the distance metric, i.e. the measure that describes how similar two samples are. Usually the simplest option, Euclidean distance, is used:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} \tag{2.11}$$

For KNN, a training data set is needed. To test how well a set of samples can be classified when their real class is known, *cross-validation* can be used. This means that one at a time each sample is classified using rest of the samples as training examples. In Bioconductor, knn1 -function and knn.cv -functions can be used to perform classification with a certain training set and by cross-validation. [52, 61]

If the classification is done between two groups (for example diseased and healthy), accuracy, precision, sensitivity and specificity of the classification can be computed by comparing the classification results to the known, true classes:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.12}$$

$$precision = \frac{TP}{TP + FP} \tag{2.13}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{2.14}$$

$$precision = \frac{TN}{TN + FP} \tag{2.15}$$

where TP = number of true positives, TN = number of true negatives, FP = false positives and FN = false negatives. In the example one has negative = healthy and positive = diseased. [61]

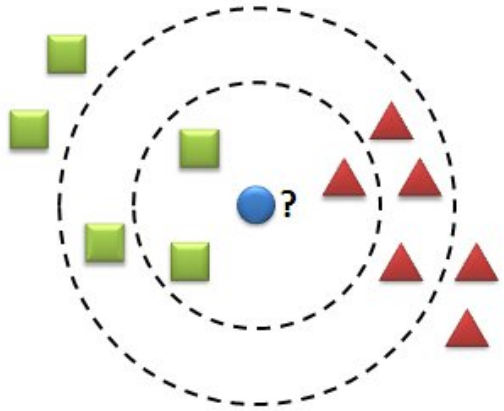


Figure 2.6: Principle of KNN classification.

2.3.3.7 PCA

Principal component analysis (PCA) is a simple, non-parametric method for extracting relevant information from data sets. PCA was invented in 1901 by Karl Pearson. The goal of PCA is to find those components or surfaces in space that show the highest variance when the data is projected to them. PCA arranges the components of the datamatrix according to their eigenvalues. [57, 61, 41]

The first principal component has the largest possible variance, and the following components have highest possible variance under the constraint that they need to be uncorrelated with the preceding components. By using only the first two or three principal components as axis in a plot and presenting the dataset as coordinate points on these axes, the data can be viewed in a sense from the most informative viewpoint. Therefore PCA is a useful tool when one wishes to visualize multidimensional data. [57, 61, 41]

PCA can be performed with a `prcomp` function from the `stats` package in R. The calculation is done by a singular value decomposition of the (centered and possibly scaled) data matrix, (not by using eigen on the covariance matrix.) [57, 70]

```
pc = prcomp(t(exprs(Data)))
```

2.4 Histopathology

Liver is very central to the metabolism of most foreign substances. The most frequent reason for the withdrawal of an approved drug from the market is stated to be drug-induced injury in liver (hepatotoxicity) [36]. Most drugs are lipophilic, which means that they can cross the membranes of intestinal cells. In the hepatocytes, they are transformed into more hydrophilic form, and may be excreted in urine or bile. This biochemical route includes oxidative pathways and transport proteins. [36]

Due to the central role of liver, toxicology studies usually include histopathological studies of liver. These studies were also performed in the Japanese Toxicogenomics Project (TGP). The most common histopathological findigs in rat liver samples in TGP were hyper-

Table 2.2: Levels of the terms used to describe the pathological findings in TGP. This grouping (blue) is used in the analysis.

• Hypertrophy
• Necrosis
• Cellular infiltration
  • Infiltration of mononuclear cells
  • Infiltration of lymphocytes
  • Microgranuloma
  • GGA
	• Vacuolar degeneration
	  • Granular degeneration
	  • Vacuolization
	  • Vacuolar degeneration
	  • Fatty degeneration
		• Cellular change
		  • Asidophilic
		  • Basophilic
		  • Eosinophilic
		  • Cellular foci

trophy, cellular infiltration and foci, microgranulomas, ground glass appearance, necrosis, degenerations and basophilic, eosinophilic and asidophilic changes (see Table 2.2). The procedure used in TGP is described in a more detailed manner in section 3.2.5. [69]

Hypertrophy is the additional development of a cell (or an entire organ) due to increase in the bulk material; the size of the cell(s) increase. **Necrosis** means the death of cell(s) in a certain area due to some external stimulus, for example due to lack of blood circulation, frost or disease. [32]

Cellular infiltration means migration and accumulation of cells within the tissues. It might also be due to excess multiplication of cells. In TG-GATEs database the histopathological findings of cellular infiltration of mononuclear cells and lymphocytes were marked separately. *Microgranulomas* are small collections of accumulated macrophages. *Ground glass appearance* (GGA) refers to uniform, finely granular cytoplasm with a peripheral clear halo. In liver, GGA occurs in the cytoplasm of liver cells due to the swelling of smooth endoplasmic reticulum. [32]

Degeneration can be determined as impairment or loss of the function and structure of cells or tissues, often leading to death of the involved part. *Vacuolar degeneration* means the formation of vacuoles in cytoplasm, most frequently due to accumulation of fat or water by cloudy swelling. *Fatty degeneration* (=steatosis, *suom. rasvamaksa*) is the accumulation of excess lipids in vesicles. In *granular (eosinophilic) degeneration*, intracellular water is accumulated in cells. *Vacuolization* means formation of any kind of vacuole in the cells. As these terms refer to similar events, in the analysis they are considered to belong into a larger group called as **vacuolar degeneration**.

Basophilic, eosinophilic or asidophilic change is a term describing the change in appearance of certain cells when staining them with different dyes. Basophilic cells take up basic dyes, and thus changes in for example nucleic acid concentrations are reflected. Eosin on the other hand is an acidic dye, and therefore changes in basic structures are seen. Asidophilic staining also uses acidic dyes. *Cellular foci* is a region of a localized bodily infection or disease, and was also used in TGP to mark cellular changes. All these findings were considered in the analysis as **cellular change**. [32]

3. METHODS AND MATERIALS

In this project, R -programming language and Bioconductor -packages were used to analyze selected microarray data available in TG-GATEs database. Some gene lists produced with Bioconductor -packages and R were also studied with Ingenuity Pathway Analysis tools. Customized CDF-files and several existing tools were used. Here the used softwares, data and analyzing tools and methods are described in detailed manner.

3.1 Softwares

The data analysis was executed in R, and several Bioconductor -packages were used as analyzing tools. The version of R used in this project is 2.15, and most of the packages used in the project are from the latest release of Bioconductor (version is 2.10). The codes needed for the analysis were written in R as separate functions. Some of the data was also analyzed using Ingenuity Pathway Analysis (IPA, Ingenuity Systems, Redwood City, CA) software.

R is a coding language for statistical computing and graphics. It is based on the S project, and was initially written by Robert Gentleman and Ross Ihaka from Statistics Department of the University of Auckland. R is available as Free Software, and can be downloaded from `http://www.r-project.org/`. [16]

Bioconductor is an open source and open development software project aiming to develop tools for the analysis of high-throughput genomic data, for example microarray data. It is based on the R programming language. The project was initiated in 2001, and it is supervised by the Fred Hutchinson Cancer Research Center. Bioconductor packages can be downloaded from `http://www.bioconductor.org`/. [20]

Ingenuity Pathway Analysis (IPA) was used to analyze and organize the gene lists of different samples which were first generated using R and Bioconductor. The knowledge of pathways, network and other interactions along with relationships between different molecules and states that were defined in the Ingenuity Knowledge Base were used widely. The Knowledge Base is "a repository of biological interactions and functional annotations created from millions of individually modeled relationships between proteins, genes, complexes, cells, tissues, metabolites, drugs, and diseases" [63]. These findings are linked to the original source of information (scientific article), manually reviewed and updated weekly. The terms and definitions are somewhat unique to the database. [64, 63]

3.2 Data

In this project we used rat liver *in vivo* gene expression data from Japanese TG-GATEs database (see 2.1.2.1 for more details) which has been publicly available since February 25, 2011. Here we describe the used data more closely.

3.2.1 Test Organisms

As mentioned in 2.1.2.1, rat was used as a model organism for the analyses in TG-GATEs. The rats were six weeks old male Sprague-Dawley rats. In each treatment group there were five rats. In the repeated dose *in vivo* study, results of which are used in this project, the rats were treated with the drug once a day for 3, 7, 14 or 28 days. For controls, the same amount of rats were treated similarly but only the vehicle was given to them. [67]

Drugs were given orally in most cases and intravenously in few cases. Vehicle used for oral route was either 0.5% Methylcellulose or corn oil, and for intravenous route it was either saline or 5/5 glucose solution. There were three different dosages: low, middle and high, mainly with ratios of 1:3:10. Usually the lowest dose corresponds to the pharmacological dose. [67]

Rats were sacrificed 24 h after the last dose, after which their body weight, liver and kidney weight, food consumption, hematology and blood biochemistry were studied. Histopathology of liver and kidney were also examined. Gene expression analysis was performed on three rats from each group. In 2006, it was estimated that data will be collected from 24 000 rats. [68, 67]

3.2.2 Compounds

The compounds used in testing were mainly well known medical compounds such as aspirin, diazepam and ibuprofen. The compounds were chosen to cover most of the therapeutic categories (such as CNS, anti-inflammatory, cardiovascular, metabolic etc.), excluding those that are not suitable for transcriptome analysis from either liver or kidney (such as dermatological drugs) and anti-cancer drugs, whose toxicity to liver or kidney is not a primary problem.

There were 150 different compounds in the database during the project. Some of the compounds were supplied by the companies that were members of the project. Withdrawn drug candidates from the member companies of the project were also studied. The compounds are listed in table 3.1. [68]

3.2.3 Procedure

Immediately after the sacrifice of the test animal, a sample of about 30 mg was taken from the left lateral lobe of the liver. The sample was stored in *RNA later* (Ambion, Austin, TX, USA) overnight at -4^oC. The RNA was isolated with RNeasy kit by Bio Robot 3000 (Qiagen, Valencia, CA, USA). Samples were homogenizated with Mill Mixer (Qiagen) and zirconium beads, and the purity of the samples was tested with gel electrophoresis. [65]

Table 3.1: The compounds studied in TG-GATEs database.

Acetaminophen	Famotidine	Perhexiline maleate
Acetazolamide	Fenofibrate	Phenacetin
Ajmaline	Fluphenazine dihydrochloride	Phenobarbital sodium
Allopurinol	Flutamide	Phenylbutazone
Allyl alcohol	Furosemide	Phenytoin
Carbamazepine	Gemfibrozil	Promethazine hydrochloride
Carbon tetrachloride	entamicin sulfate	Propylthiouracil
Carboplatin	Glibenclamide	Puromycin aminonucleoside
Cephalothin sodium	Griseofulvin	Quinidine sulfate
Chloramphenicol	Haloperidol	Ranitidine hydrochloride
Chlormadinone acetate	Hexachlorobenzene	Rifampicin
Chlormezanone	Hydroxyzine dihydrochloride	Simvastatin
Chlorpromazine hydrochloride	Ibuprofen	Sodium valproate
Chlorpropamide	Imipramine hydrochloride	Sulfasalazine
Cimetidine	Indomethacin	Sulindac
Ciprofloxacin hydrochloride	Iproniazid phosphate	Tacrine hydrochloride
Cisplatin	Isoniazid	Tamoxifen citrate
Clofibrate	Ketoconazole	Tannic acid
Clomipramine hydrochloride	Labetalol hydrochloride	Terbinafine hydrochloride
Colchicine	Lomustine	Tetracycline hydrochloride
Coumarin	Lornoxicam	Theophylline
Cyclophosphamide monohydrate	Mefenamic acid	Thioacetamide
Cyclosporine A	Meloxicam	Thioridazine hydrochloride
Danazol	Metformin hydrochloride	Ticlopidine hydrochloride
Dantrolene sodium Hemiheptahydrate	Methapyrilene hydrochloride	Tiopronin
Diazepam	Methimazole	Tolbutamide
Diclofenac sodium	Methotrexate	Triamterene
Diltiazem hydrochloride	Methyldopa	Triazolam
Disopyramide	Methyltestosterone	Trimethadione
Disulfiram	Mexiletine hydrochloride	Vancomycin hydrochloride
DL-ethionine	Monocrotaline	Vitamin A
Doxepin hydrochloride	Moxisylyte hydrochloride	WY-14,643
Doxorubicin hydrochloride	Naproxen	(7)-Chlorpheniramine maleate
D-Penicillamine	Nicotinic acid	(7)-Sulpiride
Enalapril maleate	Nifedipine	17-alpha-Ethinyl estradiol
Erythromycin ethylsuccinate	Nimesulide	2-Acetamidofluorene
Ethambutol dihydrochloride	Nitrofurantoin	2-Bromoethylamine hydrobromide
Ethanol	Nitrofurazone	
Ethionamide	N-nitrosodiethylamine	
Etoposide	N-phenylanthranilic acid	
	Omeprazole	
	Papaverine hydrochloride	
	Pemoline	

Microarray analysis was performed to 3 of the 5 samples using either rat230A or rat230.2 GeneChip probe arrays (Affymetrix, Santa Clara, CA, USA). The procedure used was based on the manufacturer's instructions. For cDNA synthesis, Superscript Choice System (Invitrogen, Carlsbad, CA, USA) and T7-$(dT)_{24}$-oligonucleotide primer (Affymetrix) were used. cDNA Cleanup Module (Affymetrix) was used for purification of the cDNA. To synthesize biotin-labeled cRNA, RNATranscript Labeling Kit (Enzo Diagnostics, Farmingdale, NY, USA) was used. [65]

Hybridization of fragmented cRNA (10 μg) to the probe array was performed at 45^o C at 60 rpm and it took 18 h. After hybridization, the arrays were first washed and then stained with streptavidin-phycoerythrin (Fluidics Station 400 by Affymetrix was used). For scanning, Gene Array Scanner by Affymetrix was used. For image processing, Affymetrix Microarray Suite version 5.0 (MAS 5.0) was used. Per chip normalization was performed by setting the mean intensity to 500. [65]

3.2.4 Affymetrix GeneChip: Rat Genome 230 2.0 Array

The general principles of Affymetrix GeneChip arrays are presented in 2.2.3.1. Affymetrix GeneChip Rat Genome 230 2.0 Array is a whole-genome array that has 31,000 probe sets that can be used to analyze expression of over 30,000 transcripts from over 28,000 well tested genes. Sequences used to design the array were selected from GenBank, dbEST and RefSeq. The sequence clusters were created based on UniGene database, and developed further by comparing them to the draft assembly of rat genome from the Baylor College of Medicine Human Genome Sequencing Center. [2]

The oligonucleotide probes complementary to each sequence are synthesized to the array *in situ*. Each sequence is represented by 11 pairs of oligonucleotide probes. A set of rat maintenance genes are included in the array for normalization and scaling of the experiments. [2]

3.2.5 Histopathological Analysis

After the sacrification of the rats, their livers were quickly removed, and sections of them were placed to 10% phosphate-buffered formalin. These formalin-fixed sections were then embedded in paraffin, and stained witht hematoxylin and eosin. Then the samples were examined with light microscopy. [39, 67]

Each sample was analyzed by two trained pathologists. They graded the findings to no change, minimal, slight, moderate and severe. In figure 3.1 for example periportal necrosis was detected. [39, 67]

3.2.6 Blood Biochemistry

Several concentrations were measured from the blood samples (collected from the abdominal aorta). These include: alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), total bilirubin (TBIL) and Lactate dehydrogenase

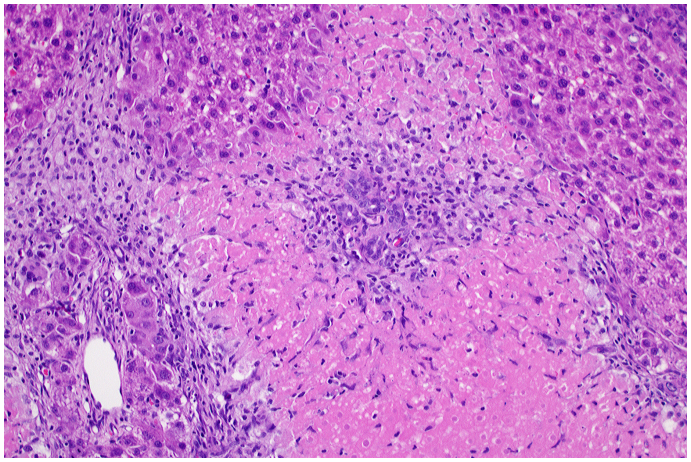


Figure 3.1: Example of histopathological sample from TG-GATEs. Slight periportal necrosis was detected from this sample of a rat that had been given 30mg/kg of allyl alcohol for 15 days.

(LDH). All these values are used in medical examination to predict liver damage or malfunction. [22]

Alanine transaminase (ALT), also called serum glutamic pyruvate transaminase (SGPT) or alanine aminotransferase (ALAT) is an enzyme present in hepatocytes. Aspartate transaminase (AST), also called serum glutamic oxaloacetic transaminase (SGOT) or aspartate aminotransferase (ASAT), is similar to ALT as it is another enzyme associated with hepatocytes. AST value is not that specific to liver damage. [22]

Alkaline phosphatase (ALP) is a protein found in all tissues, but especially in high levels in the liver, bile ducts, and bone. ALP is a hydrolase enzyme which removes phosphate groups from many types of molecules (such as nucleotides and proteins). Lactate dehydrogenase (LDH) is an enzyme found in many body tissues, including the liver. Bilirubin is a yellowish pigment found in bile, which liver helps to break down. [22]

3.2.7 Arrangement of the Data into Histopathological Groups

For the analysis, the data was grouped based on the histopathological findings from each array. For this purpose we used a table with all the necessary information about the arrays reported in TG-GATEs. The table contained following information about each array: the barcode, the batch number, used chemical, dose, dose level (control, low, middle, high), exposure time (3, 7, 14, 28) and the most severe histopathological finding and its grade (minimal, slight, moderate, severe). If there were several findings with equal severity, alphabetical order was used to decide which term to use.

Based on this, we chose few of the most common findings as described in Chapter 2.4, and grouped the files according to them. Table 3.2 shows the number of samples found in each group. All the available tests done with vehicles were used as controls.

To analyze the dose response, the data was arranged a bit differently: first, those samples with a certain histopathological finding and treated with the highest dose were

Table 3.2: The most common histopathological findings and amount of samples in each group.

(Histopathological) group	Number of samples
Hypertrophy	442
Necrosis	77
Cellular change	124
Cellular infiltration	344
Vacuolar degeneration	173
Controls (without findings)	1392
Treated, but no phenotypical finding	3621
Other findings	591
Total number of samples	6765

chosen for each group ("high dose" -group). Next, the compounds found in each group, i.e. causing a certain phenotype, were determined. Then the compounds which caused this certain phenotype in more than five samples were chosen (see Fig. 4.12), and those samples treated with the same compound but with middle or low dose, and which show nothing in the phenotype (=no pathological findings reported), were chosen as the other two groups ("middle dose" and "low dose").

The deviation of different compounds causing different phenotypes can be seen in Fig. 4.12. Note that for example necrosis is not caused by any particular compounds alone, but is found from many groups. In all the cases, there were some findings in controls as well, but these were not used for the analysis.

Similar analysis was attempted for the different exposure times using only the highest dose, but there was enough data only if the two latest time points (28 and 14 days) and the two earliest time points (3 and 7 days) were combined. Even with these conditions, only one group (hypertrophy) contained enough data for reliable analysis.

3.3 Analysis Methods

As mentioned in 2.3, there are several options for each step of microarray data analysis. In the same chapter the principles of different methods were briefly discussed. Here we describe which methods were used in this project to analyze the previously described data.

3.3.1 Preprocessing

Before the actual data analysis steps, the data was first loaded to R and preprocessed, and batch effect correction and some filtering steps were performed. Here these procedures are described more closely. Also the Chip Description Files (CDF) used are presented here.

3.3.1.1 Preprocessing: justRMA

The data was downloaded from TG-GATEs as .CEL -files, which contain one single representative intensity value for each feature [13]. There were 6765 of these files (32.6 GB),

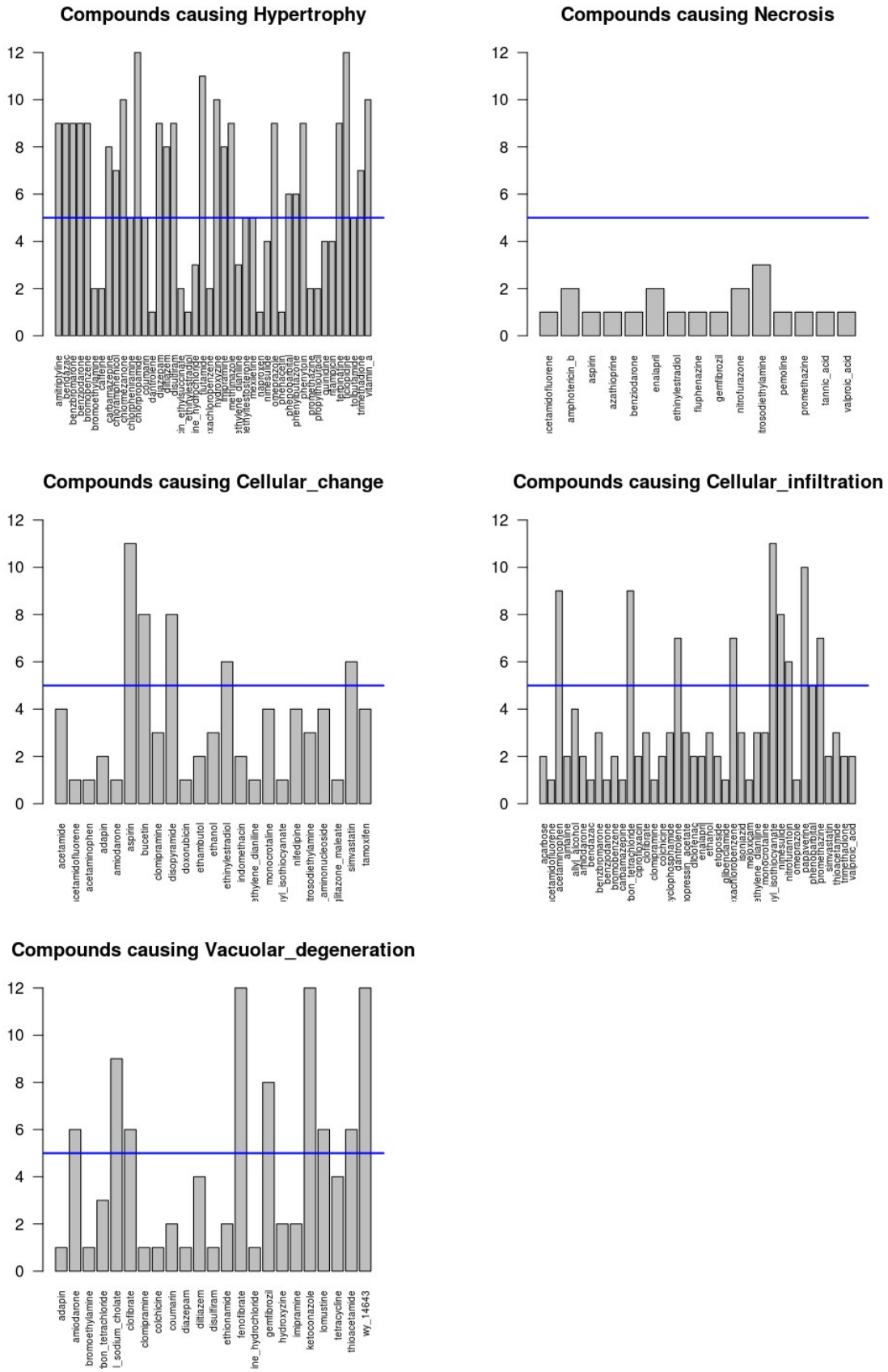


Figure 3.2: Different compounds causing certain histopathological findings with high dose. For further analysis, only those compounds causing certain finding in more than five samples were considered (blue line).

so an effective function was needed to load them to R for the analysis. Thus a function called `justRMA` was used to read the CEL files and to simultaneosly perform preprocessing of the files. It can be said that `justRMA` combines `ReadAffy` and `rma` -fucntions. [18, 51]

The `justRMA` -function performs normalization using quantile normalization. For background correction it uses RMA correction. Both methods are described in section 2.3.3.2. The CDF-files were given as a parameter to the function. Output of the function is an `ExpressionSet` -object. [18, 51]

3.3.1.2 Filtering

After normalization of the data, non-specific filtering was performed to reduce the amount of the data and to make the computations less time-consuming. Those probesets that show little variation across all the samples were filtered out, as they do not contain any valuable information. First, the coefficient of variability (CV = standard deviation / mean) were computed for each probeset. By using CV -values instead of just standard deviations makes filtering less sensitive to intensity related bias. It was chosen to include 10% of the genes with highest variabilty (see Figure 3.3). With `nsFilter` function, control probes were filtered using the `AFFX` - prefix in their name.

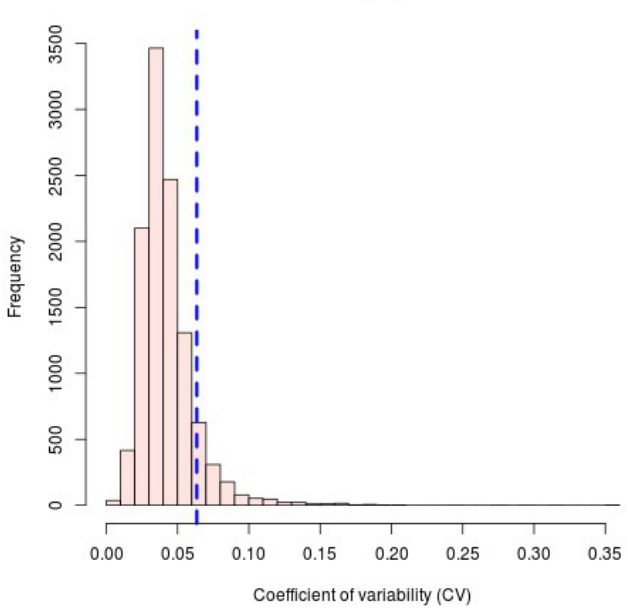


Figure 3.3: Histogram showing the filtering according to coefficients of variability. Blue line shows the 10% cut-off used for filtering.

3.3.1.3 Batch Effect Correction

In the data used in this project, the batches usually contained some or all of the experiments with same compound and same exposure time (that is, for example all 3 days experiments

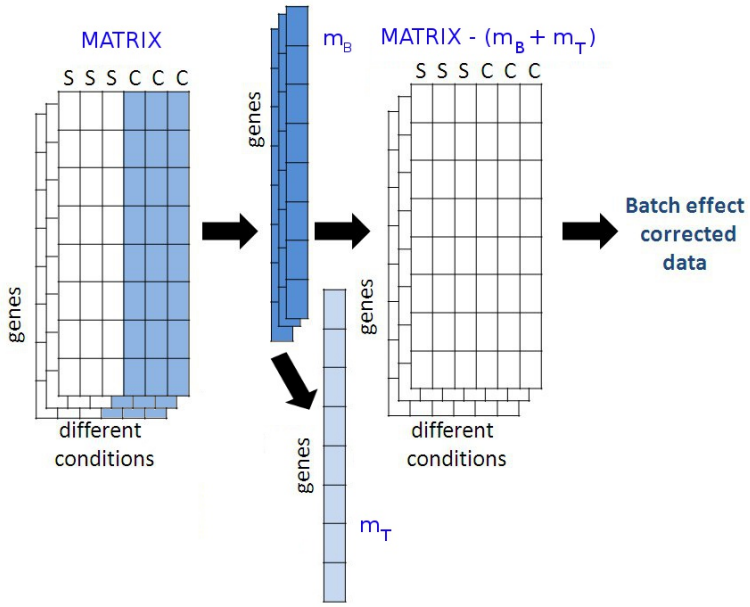


Figure 3.4: Illustration of the principle of the batch effect correction used in this project.

done with aspirin with different doses, including the controls). Because we organized the data based on the histopathological findings and thus mixed together files from experiments with different compounds, exposure times and dosages, a special batch effect correction was needed. For this purpose, a special R-script was written.

The idea behind the batch effect correction script is that all the gene values in each control group should in theory be at the same level (lets call this the *zero level*): an assumption is made that all the variation between the different controls is due to the batch effect. Hence the values of the control experiments can be used to correct the batch effect in all the arrays.

The script uses the information from the TG-GATEs [47] described in section 3.2.7. It computes the average of each gene value of all the control arrays for a given batch i (m_B). This value is then subtracted from each corresponding gene in each array that belongs to the batch i in order to set the zero level of each batch to the same level. As this method shifts the zero level to zero, we also calculate a universal mean of each gene as a mean of all the controls (m_T), and add this value correspondingly to each gene of all the arrays. See the illustration in Fig. 3.4.

3.3.1.4 Custom CDF -files

As stated before, in Affymetrix GeneChips a probe set contains 11-20 pairs of oligonucleotides to represent one target gene or transcript. However these probes are based on earlier genome annotations, which are significantly different from current knowledge due to the tremendous progress in genome sequencing in recent years. This new information of gene sequences, directions and alternative splicing causes several problems and affects the analysis results. [12]

To solve this problem, Dai et al [12] have developed a procedure for reorganizing the

Table 3.3: The principle of the design matrix used to fit linear models to data.

File	Control	Disease
Vehicle1.CEL	1	0
Vehicle2.CEL	1	0
Hypertrophy1.CEL	0	1
Hypertrophy2.CEL	0	1
Hypertrophy3.CEL	0	1
Vehicle3.CEL	1	0

GeneChip probes into probe sets. The procedure is described in [12], but in short it performs sequence alignments of GeneChip probe sequences with the most recent genome sequence of the species in question (using BLAST), and chooses only those probes to a probe set which have only one perfect match to the genome sequence and certain cluster. It is also required that each probeset contains at least three probes and that they are all from the same strand and have same direction. This procedure alters the sizes of probe sets from three to several dozens of probes. [12]

The re-organized probe sets can be included in the data analysis by using Chip Description Files (CDF). A CDF -file describes the link between probes and probesets, and identifies PM and MM probes and control probes. A new CDF-file is usually generated few times a year, and they can be downloaded from the website of the Microarray lab of Molecular and Behavioral Neuroscience Institute of University of Michigan: (http://brainarray.mbni.med.umich.edu/brainarray/Database/CustomCDF/genomic _curated_ CDF.asp). Use of these customized CDF-files can improve the analysis results by more than 20% (20% improvement was gained with HG-U133A -chip). Custom CDF-files can be used in Bioconductor according to this example:

```
data <- ReadAffy(cdfname= ''HS133A_HS_UG'')
```
[12]

Nowadays also Affymetrix annotation system maps probe sets to the latest UniGene (http://www.ncbi.nlm.nih.gov/unigene) build every couple of months. Nevertheless it does not react to situations where some of the probes in a probe set are assigned to another gene or to multiple genes. Also many of the UniGene clusters are represented by multiple probe sets in GeneChips. The custom CDF-files correct these issues as well. [45]

In this project we used Version 14 (released April 2nd, 2011) ENSG -files for rat 230.2 Affymetrix arrays. Genome wide annotation-file for rat (`org.Rn.eg.db`) from Bioconductor was used [9].

3.3.2 Differentially Expressed Genes

To analyze differential expression, linear models and `limma` -package for R were used. For this purpose, a design -matrix was build, showing which samples are controls (treated only with the vehicles) and which have some histopathological finding (i.e. disease). Table 3.3 demonstrates how the design matrix was organized. [24, 60]

After this, `lmFit` -function was used to fit a linear model for every gene. Next, a

contrast matrix was created from the design matrix using `makeContrasts` -function. Then the coefficients and standard errors were computed with `contrasts.fit` -function. Finally, standard errors were moderated using empirical Bayes method and `eBayes` -function. [24, 60, 59]

With `topTable` -function genes can be organized based on their p-values. The differentially expressed genes were picked up from this list using cut-offs for p-values and for fold-changes. In MicroArray Quality Control (MAQC) project [56] they present a recommendation, that for a gene to be considered as differentially expressed, the p-value should be less or equal to 0.001 and the mean difference greater than or equal to twofold. These values are good for smaller datasets, but to see enough differential expression in this large dataset, a p-value cut-off of 0.05 and fold-change cut-off of 1.5 were used. [24, 60, 59]

3.3.3 GO Enrichment

After determining the the differentially expressed genes, an attempt was made to identify whether these genes form some gene sets: that is, whether some gene set is under- or overpresented in the data. For this purpose function called `hyperGTest` was used. To perform this hypergeometric testing, a gene universe and a list of interesting genes are needed. Gene Ontologies (GO) were used to determine the gene sets. [24]

First, a *gene universe* was defined as all the possible unique genes that exist on the rat 230.2 array. *Ensembl* gene identifiers from the custom CDF-file were used to create the universe. The interesting genes were the previously determined differentially expressed genes. In addition to these, the name of the annotation data package (`org.Rn.eg.db`), the GO ontology of interest (`BP`, Biological Process), the p-value cut-off (`0.05`) and the test direction (`over` for overpresented) were given as parameters to `hyperGTest` -function. [24, 21]

The results of the analysis were saved in a html-file using `htmlReport` -function. This allows closer examination of the GO term names via AmiGO -gene ontology webpage [48, 8, 21], as the names of GO terms are printed as html links. [24]

3.3.4 PCA

The principal component analysis (PCA) was performed to visualize the separation of the controls and the arrays with histopathological findings. PCA was performed separately to each histopathological group versus the controls.

Function `prcomp` was used to compute the principal components. Only those 50 probes showing lowest p-values were used in the analysis. Expression matrix of these 50 probes was transposed so that the function understood the gene expression levels as features of the sample objects. The code is presented in Appendix A, Chapter A.1.

3.3.5 KNN

K-nearest neighbor (KNN) classification by cross-validation was used to determine how well each histopathological group can be separated: for each group, one sample at a time were classified using rest of the samples (controls and other samples from that group) as training examples. Three closest neighbors were used for classification. For this purpose, the gene expression values were first standardized (as is recommended in [24], see Chapter 9, Supervised Machine Learning). Then the Euclidean distances are computed (using `dist` function). After this, the differentially expressed genes were organized according to their t-statistic, and 50 first ones were used for the analysis. Finally, `knn.cv` function was used to compute the classification results.

These results were then compared to the real classes (that is, was the sample marked as control or to have certain histopathological findings). Accuracy, precision, sensitivity and specificity of each classification test was computed. The code is presented in Appendix A, Chapter A.2.

3.3.6 Ingenuity Pathway Analysis

After the genes were listed and organized using the `limma` -package as described in section 3.3.2, these gene lists were uploaded to Ingenuity. For this purpose, an excel sheet was produced with three columns: gene ID annotations, log fold changes and adjusted p-values. Suitable cut-offs were chosen, usually 0.001 for p-value and 1.2 for the fold change (equals to 0.26 for log fold change). Both up- and down-regulated genes were considered.

"Core Analysis" function analyzes the data in the context of biological processes, pathways and networks. Significance of the different groups is tested using Fishers' Exact test p-value. Some (Core, Tox or Metabolomics) "Comparison Analyses" were also performed compare different groups or to compare it to data from some single samples treated with certain drugs. Lots of data from TG-GATEs and some using new possible drug candidates were uploaded to Ingenuity earlier.

4. RESULTS

In this chapter, analysis results are presented and briefly discussed. First, the lists of differentially expressed genes (DEG) and the enriched gene ontology (GO) categories amongst them are presented. To examine these DEGs and GOs further, the gene lists were uploaded to Ingenuity: some results from Core and Comparison Analyses are presented here. Next the PCA and KNN results of separation between the five histopathology groups from healthy controls are discussed. Correlation to other phenotypical findings are presented. Those samples without any reported histopathological findings are examined further, and some dose and time response analyses using these samples are shown.

Finally, the data from the five groups is compared to two known drug molecules and two possible drug candidates. Analysis results from similar comparison analysis using different dose levels is also shown. The gene expression data from each group is also compared to known toxicity -related pathway. Compounds responsible for the histopathological findings are also briefly mentioned and discussed.

4.1 Differentially Expressed Genes and GOs (Biomarkers)

For each of the five histopathology groups, differentially expressed genes (see Table 4.1) and enriched gene ontologies (GO) among these were listed. In Table A.5 (Appendix) one can see as an example the most differently expressed genes from comparison between cellular change -group and controls. Gene symbols and names are presented together with $\log_2$ fold-changes and adjusted p-values. For example genes Cyp2f4 (cytochrome P450) and Apex1 (APEX nuclease) are reported to be regulated by toxic compounds [72]. To get a better idea of what the changes in expression of these genes mean, GO enrichment was performed. The gene lists were also uploaded to Ingenuity IPA for pathway analysis (Chapter 4.2).

In Table A.10 (Appendix) the 50 enriched up-regulated biological process GO -categories with lowest p-values are presented. One can see that many GO terms implying exposure to interesting substances are amongst the enriched categories, for example "response to drug", "response to toxin", "cellular response to chemical stimulus" and "response to stress". Several terms seem to be related to lipid metabolism (for example "lipid metabolic process", "fatty acid metabolic process" and "lipid biosynthetic process"). Terms like "response to nutrient levels" and "response to nutrient" might indicate the effects of the new diet including the corn oil used as the vehicle when giving the compound as oral dose. These issues were examined further with Ingenuity Core Analysis and Comparison Analysis.

Table 4.1: Number of the samples and differentially expressed genes (DEGs) in each histopathology group (cut-offs used: adjusted p-value < 0.001 and log fold-change 1.2). Also the number of different compounds causing certain histopathological finding and the number of those compounds that cause this finding in more than five treated samples are presented.

Histopathology group	No. of samples	No. of DEGs	No. compounds (all/in $>$ 5 cases)
Hypertrophy	442	229	47/28
Necrosis	77	66	40/1
Cellular change	124	98	26/8
Cellular infiltration	344	197	52/23
Vacuolar degeneration	173	880	25/ 10

4.2 Core Analysis Results

Interesting genes from each group were uploaded to Ingenuity IPA, and Core Analysis was performed. Some results of these analysis are presented in Table 4.2, and in Appendixes. From each analysis, Top Bio Functions of Molecular and Cellular Functions, Canonical Pathways, Upstream regulators, Tox Lists, Tox Functions for Hepatotoxicity and Networks are presented in the tables. Some of these results are also visualized in Fig. 4.1.

In Top Bio Functions and Top Tox Functions, the number of genes or molecules is presented in parenthesis. In Canonical Pathways and Top Tox Lists the ratio of molecules in the pathway meeting the cutoff -criteria and the total number of molecules that make up that pathway is given. In Top Networks, a network score is presented in parenthesis: the score represents the likelihood that the molecules that are part of the network are found by random chance alone. Mathematically it is the negative exponent of right-tailed Fishers exact test result: for example score of 3 means that there is a 1 in 1000 chance that the molecule is found by chance. [64]

In Top Bio Functions (Molecular and Cellular Functions) three groups arise in all the five cases: Lipid Metabolism, Molecular Transport and Small Molecule Biochemistry. Lipid Metabolism is related to both steatosis and cholestasis (via cholesterol and bile acid biosynthesis). Molecular transport means functions associated with intra- and extracellular movement of any kind of molecules, and it may reflect the increased expression and movements caused by it. Small Molecule Biochemistry describes functions associated with small molecules such as nitric oxide or indole: activity of these functions may reflect the treatment with the drugs.

Same things are visible also in Canonical Pathways; ”fatty acid metabolism” and ”glycerolipid metabolism” reflect steatosis, and pathways like ”Xenobiotic Metabolism Signaling” and ”Metabolism of Xenobiotics by Cytochrome P450” indicate the treatment with drugs. Also ”LPS/IL-1 Mediated Inhibition of RXR Function” and ”PXR/RXR Activation” pathways are seen in many cases: these pathways are related to hepatic acute phase response (APR), which can be caused by infection, inflammation or injury. Very similar lists are

Table 4.2: Top biofunctions, canonical pathways, upstream regulators, tox lists and networks from Ingenuitys Core analysis of differentially expressed genes from those samples with **hypertrophy** as histopathological finding. In Top Bio Functions and Top Tox Functions, the number of genes or molecules are presented in the parenthesis. In Canonical Pathways and Top Tox Lists the ratio molecules in that pathway that meet the cut-off criteria and the total number of molecules that make up that pathway is given. In Top Networks a network score representing the likelihood that the molecules are found by change is shown.

Hypertrophy-group (229 differentially expressed genes)	
Top Bio Function, Molecular and Cellular Functions	**Top canonical pathways**
Lipid Metabolism (67)	Metabolism of Xenobiotics by Cytochrome P450 (12/196)
Small Molecule Biochemistry (83)	LPS/IL-1 Mediated Inhibition of RXR Function (17/236)
Molecular Transport (64)	Xenobiotic Metabolism Signaling (18/295)
Drug Metabolism (28)	Fatty Acid Metabolism (12/183)
Energy Production (23)	Nitrogen Metabolism (7/119)
Top Upstream regulators	**Top Tox Lists**
PPARA (Activated)	Xenobiotic Metabolism Signaling (23/347)
ciprofibrate (Activated)	LPS/IL-1 Mediated Inhibition of RXR Function (19/246)
POR	NRF2-mediated Oxidative Stress Response (18/231)
pirinixic acid (Activated)	Fatty Acid Metabolism (13/123)
ACOX1 (Inhibited)	CAR/RXR Activation (7/29)
Top Tox Functions, Hepatotoxicity	
Liver Steatosis (19)	
Hepatocellular Carcinoma (21)	
Liver Hyperplasia/Hyperproliferation (21)	
Glutathione Depletion In Liver (5)	
Liver Cholestasis (6)	
Top Networks	
Drug Metabolism, Glutathione Depletion In Liver, Carbohydrate Metabolism (47)	
Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry (37)	
Cellular Assembly and Organization, DNA Replic., Recomb. and Repair, Molecular Transport (33)	
Lipid Metabolism, Small Molecule Biochemistry, Vitamin and Mineral Metabolism (32)	
Ophthalmic Disease, Respiratory Disease, Hereditary Disorder (28)	

found from the Top Tox Lists as well.

Steatosis and Cholestasis emergence also in Hepatotoxic Top Tox Functions. Hepatocellular Carcinoma, Liver Hyperplasia/Hyperproliferation and Glutathione Depletion are also seen in many cases, which makes sense, even though one would have expected to see highest activity of steatosis and cholestasis in vacuolar degeneration -group and hyperplasia/proliferation in hypertrophy. It should also be considered that even in hypertrophy-group the bar for hypertrophy Tox Function is not over the threshold line: this might reflect the differences in determination of terms in TGP and in Ingenuity, but also the fact that Ingenuity database is created by text-mining scientific papers and it includes all the data available from different organisms, organs and methods.

Top Networks seem to reflect the necrotic situation in cells with histopathological issues (Cell Morphology, Organ Development, Cell Cycle, Cell Death, Cellular Assembly and

Organization), steatosis (Lipid Metabolism), cholestasis (Endocrine System Development and Function, Nutritional Disease) and oxidative stress (glutathione depletion). Of the upstream regulators, PPARA (peroxisome proliferator-activated receptor alpha) and POR (P450 (cytochrome) oxidoreductase) come up in most cases. Activity of both molecules is known to correlate with steatosis.

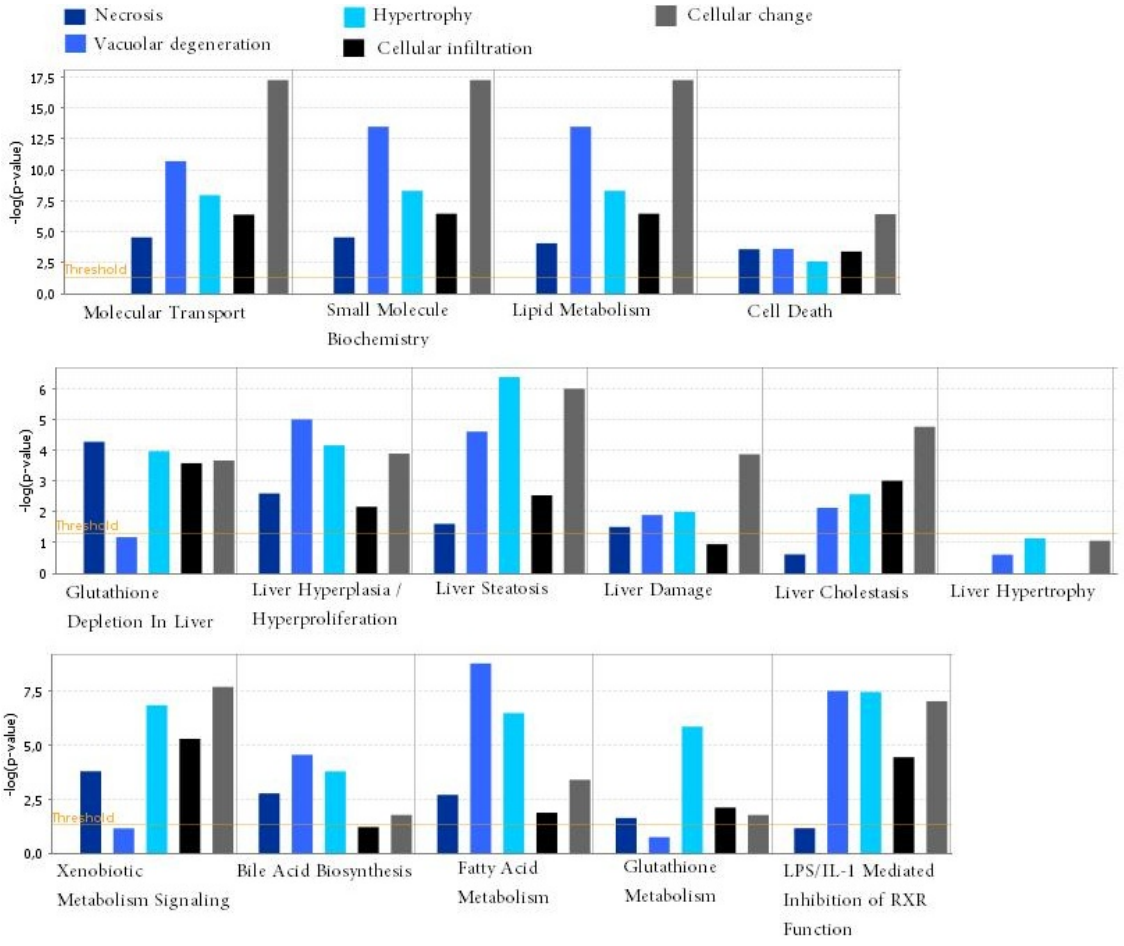


Figure 4.1: Some results of the Comparison Analysis done using Ingenuity IPA. From top to bottom: Top Bio Functions, Top Tox Functions and Canonical Pathways.

4.3 PCA Results

Principal component analysis was performed for each of the histopathological groups and controls. In (a) in Fig. 4.2, one of these plots is presented: controls are plotted with green and samples with a certain histopathological finding in blue. More figures for rest of the groups can be found from the Appendixes. Some groups form tight clusters whereas others are more spread.

To get some explanation to the deviation of the data points in PCA plots presented above, points were visualized with different colors according to the dose, exposure time and degree of the pathological finding. Some examples of these plots are shown in Fig. 4.2, rest of them can be found from the Appendixes. From these plots one can see that

in some cases the doses, exposure times and severities of the finding seem to explain the deviation (see for example hypertrophy -group).

4.4 KNN Results

KNN cross-validation (k=3) results are presented in Table 4.3 and are used to understand how well different groups are separated. Precision, accuracy, sensitivity and specificity were computed for each of these classifications: these results are presented in Table 4.4. The results are in agreement with the PCA results: all the groups detect the controls as one group nicely (good specificity), but in all groups some percentage of the diseased samples are classified as controls (not that good sensitivity).

As can be seen from the PCA images (Fig. A.1), vacuolar degeneration is most easily separated (86 % of the samples are correctly classified). Rest of the results show surprising sensitivities when compared to PCA: one would expect perhaps a lower percent for cellular change and hypertrophy (75 % and 82 %, respectively), and higher for necrosis, of which only 39 % of the diseased samples were correctly classified.

Table 4.3: Results of KNN cross-validation analysis (k=3) for all the five groups. In each case, the column on the left represents the real groups, while the rows are the classified groups. For example for necrosis: all the controls were classified correctly, while 47 of those samples with necrosis were classified as controls, and 30 samples with necrosis were classified correctly. 50 most differentially expressed genes were used for the analysis.

	control	Necrosis
control	1392	0
Necrosis	47	30
	control	Vacuolar degeneration
control	1392	0
Vacuolar degeneration	24	149
	control	Hypertrophy
control	1392	0
Hypertrophy	78	364
	control	Cellular infiltration
control	1392	0
Cellular infiltration	157	187
	control	Cellular change
control	1392	0
Cellular change	31	93

KNN cross-validation was also performed for all the five groups and controls. For this analysis, all the gene values were used. The results are shown in Table 4.5.

Again, controls are classified nicely as their own group, but there are altogether 595 samples ($\approx$ 50%) that are misclassified as controls. The situation is worst with cellular infiltration and necrosis -groups as expected. The results are generally "worse" than in

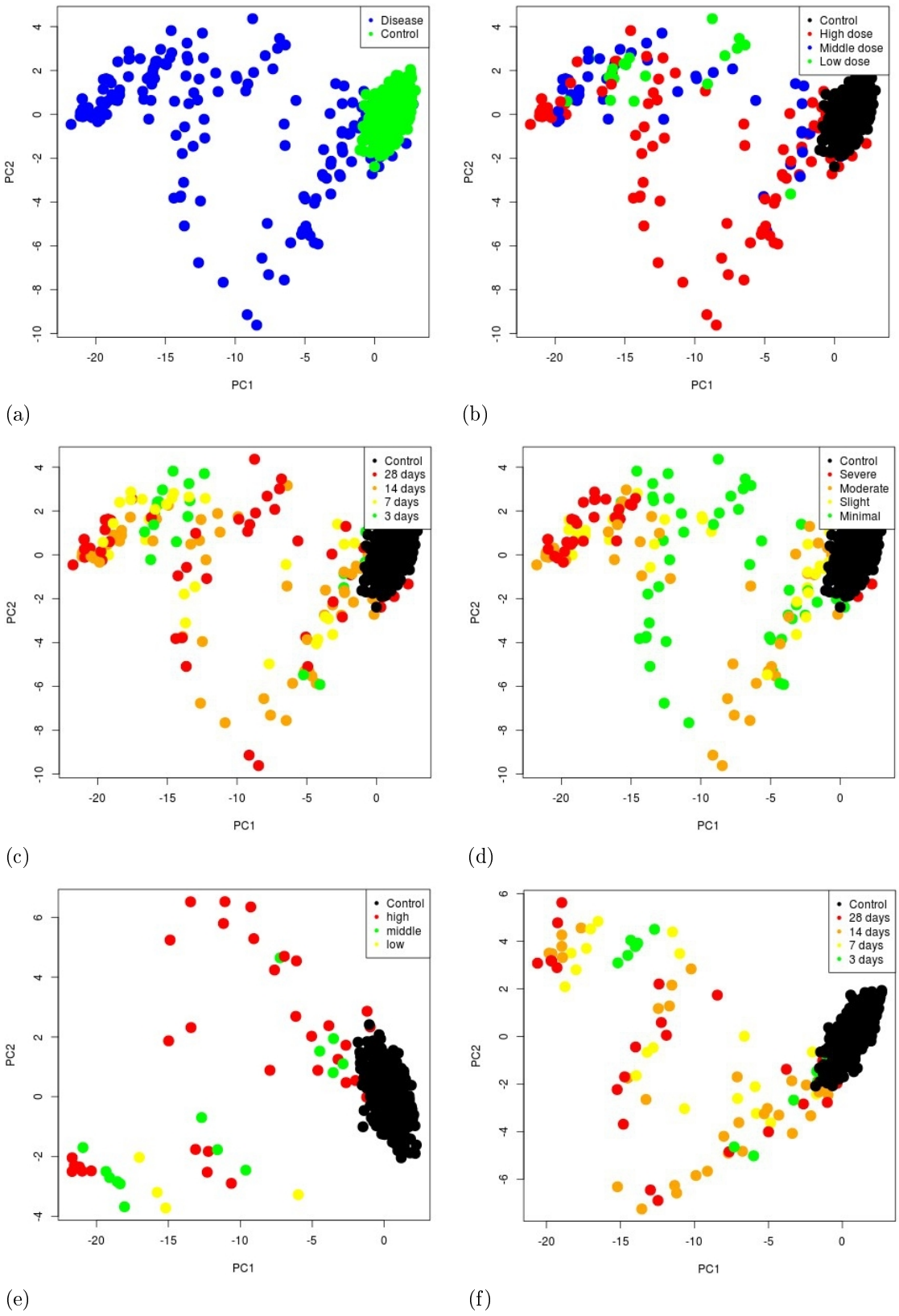


Figure 4.2: A PCA plot showing the separation between vacuolar degeneration and controls (a), also with different dose levels (b), exposure times (c) and severities of the findings (d) presented with different colors. Also different doses in a single time point (14 days) is shown (e as well as different time points in one dose level (f).

Table 4.4: Precision, accuracy, sensitivity and specificity of each group vs. controls cross-validation test with KNN.

Group:	Precision	Accuracy	Sensitivity	Specificity
Hypertrophy	1,00	0,96	0,82	1,00
Necrosis	1,00	0,97	0,39	1,00
Cellular change	1,00	0,98	0,75	1,00
Cellular infiltration	1,00	0,91	0,54	1,00
Vacuolar degeneration	1,00	0,98	0,86	1,00

case-wise comparison (Table 4.3), which is expected due to use of several groups and use of all the gene expression values instead of just the 50 most differentially expressed.

Table 4.5 can be used to study the similarities between the histopathology groups. For example, hypertrophy and cellular infiltration -groups seem to be similar at least to some extend (about 5 % missclassifications).

Table 4.5: KNN cross-validation (k=3) results of all groups. Rows present the real classes and columns the classes based on KNN.

	Cellular change	Cellular infiltration	Control	Hypertrophy	Necrosis	Vacuolar degeneration
Cellular change	**51**	7	58	3	1	4
Cellular infiltration	3	**99**	216	20	1	5
Control	1	0	**1391**	0	0	0
Hypertrophy	1	18	219	**201**	1	2
Necrosis	3	2	56	1	**10**	5
Vacuolar degeneration	2	7	46	7	0	**111**

4.5 Deviation of the Gene Expression Values

To visualize and study the actual difference in the deviation of the gene expression values between controls and the histopathological groups, boxplots were generated. The highest possible deviation in normal situation was also studied. The batch corrected and linearized data of the most differentially expressed genes were used for these plots. For the example plots presented in Fig. 4.3, ten of the most differentially expressed genes are shown for hypertrophy and necrosis -groups. Example boxplots for the rest of the histopathology groups can be found in Appendices.

From Fig. 4.3 one can see that in case of hypertrophy (a) the deviation of the gene expression values between controls and samples with hypertrophy findings differs nicely in all ten genes. The difference is greatest with *Utg1a1* and *Gstm4* genes, but even with these genes some overlap is seen in the whiskers and some outliers exist. However, this plot encourages the study of certain gene expression values as possible biomarkers.

With necrosis the situation is different, as can be seen from boxplot (b) in Fig. 4.3. The boxes of controls and treated samples overlap in all the ten cases, and the mean values (50th percentiles, the line in the middle of the box) are not very far from each other. Compared to hypertrophy group, the boxes and whiskers are much wider in necrosis (compare for example *Gsta5* -gene shown in both plots). This reflects the variety of mechanisms causing necrosis, and the result is in agreement with KNN and PCA -results.

These plots give an idea of the deviation of gene expression values in normal situation as well: in (a) and (b) in Fig. 4.3 the gene expression values in controls vary only slightly, even though there were as many as 1392 control samples. Higher deviation in the treated samples is explained partly by the different doses and exposure times. To see the highest possible deviation of gene expression values in normal situation, those 10 genes with highest variation in the controls were chosen and plotted (Fig. 4.4). These genes take care of normal daily variation in cells, and due to their high variation they are seldom listed as differentially expressed genes and could not be used as biomarkers.

4.6 Correlation Between Histopathological Findings and Other Phenotypic Findings

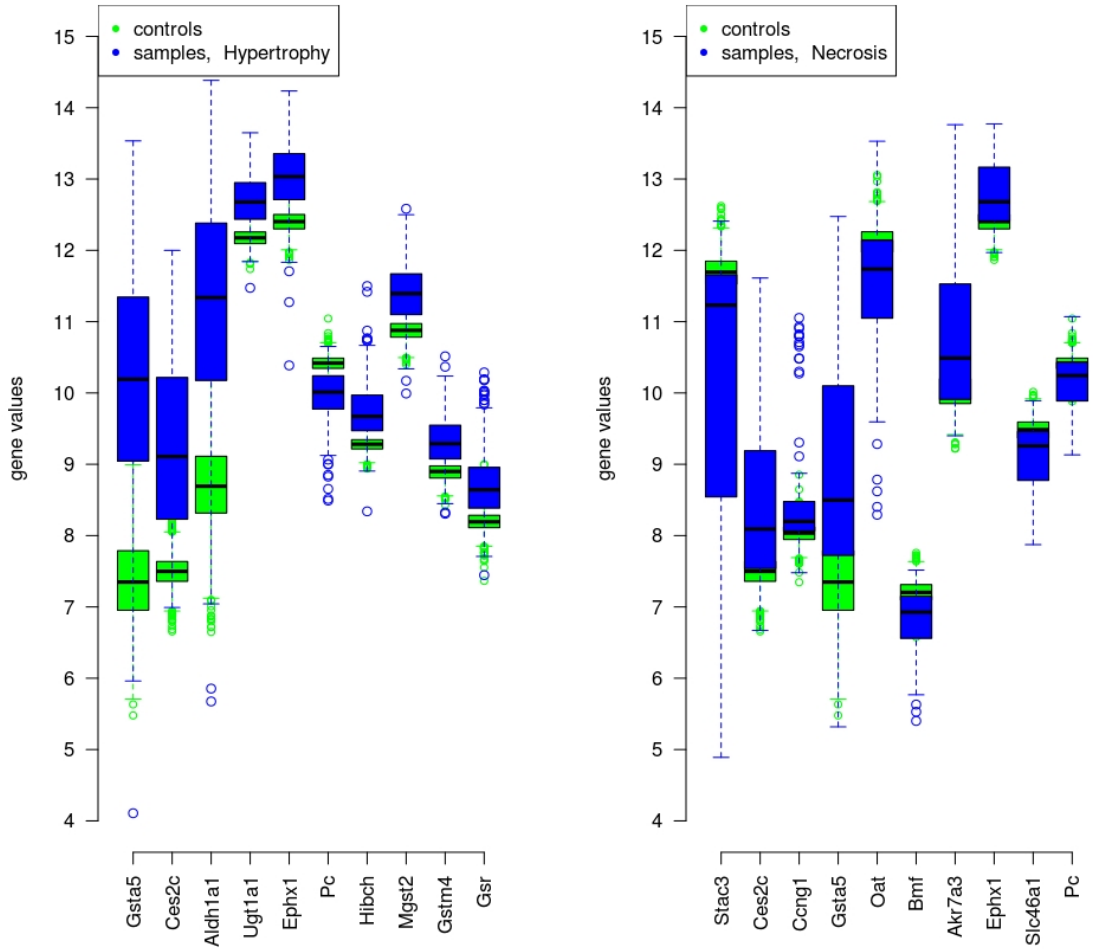
In Figure 4.5 the correlations between the histopathological groups and some other phenotypical findings are shown. For comparison, controls are shown on the left in each plot. Chosen phenotypical findings are: (from top to bottom, right column:) ALT, AST, ALP and (from top to bottom, left column:) TBIL, LDH and weight loss per day (see Chapter 3.2.6 for more information about these blood biochemistry values).

Figure 4.5 shows that there is relatively little correlation between a certain histopathological finding and any of the other phenotypical findings. Controls seem to show quite wide variation in all the findings. Perhaps these results point out that even though the blood samples in human indicate liver damage quite well, in rats the situation is different. The only signs of correlation can be seen with weight loss: all control rats have gained weight or their weight has remained the same, while in all histopathology groups some samples have lost weight.

To take a closer look on the deviation of the biochemistry values and weight loss, histograms were used to compare the values to controls. As an example the weight loss histogram is shown in Fig. 4.6. It can be seen that the mean of the weight loss is slightly shifted to right in most cases.

4.7 Treated Samples Showing No Phenotypical Findings

In the data there were large amount of samples treated with some compound but from which no phenotypical findings were reported (3621 samples, see Table 3.2). Also their spread compared to the controls was studied to see how much they resemble the controls. In Fig. 4.7 the produced PCA plot can be seen on the left. Clearly most of the samples are very close to the controls, which is expected as they might be treated for only three



(a) (b)

Figure 4.3: Boxplots showing deviation of gene expression values of chosen genes in controls (green) and samples (blue) belonging to hypertrophy (a) and necrosis (b) groups. (The bottom and top of the box are the 25th and 75th percentile, the band near the middle of the box is the 50th percentile, and the whiskers show the most extreme data points which are no more than 1.5 times the interquartile range from the box. Data points beyond this range (outliers) are marked with circles.

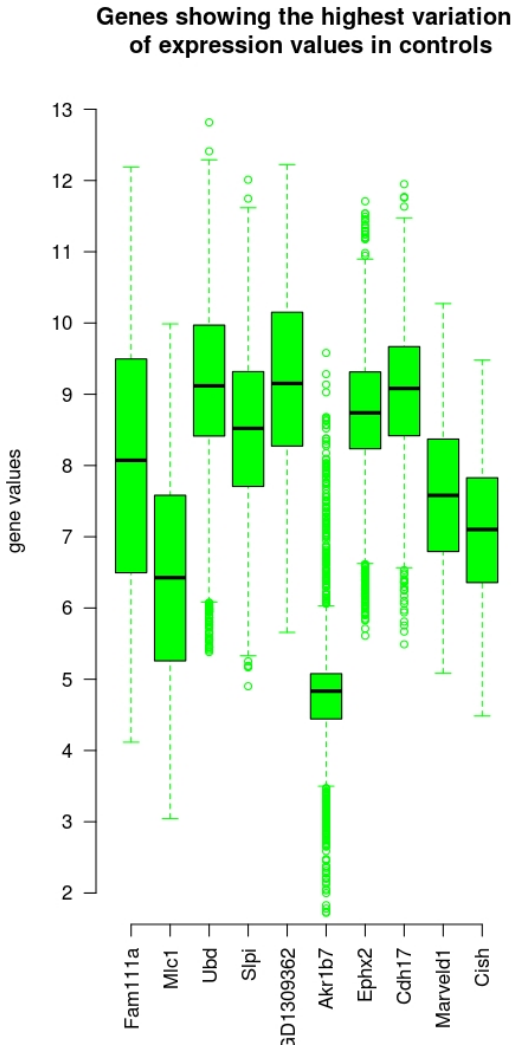
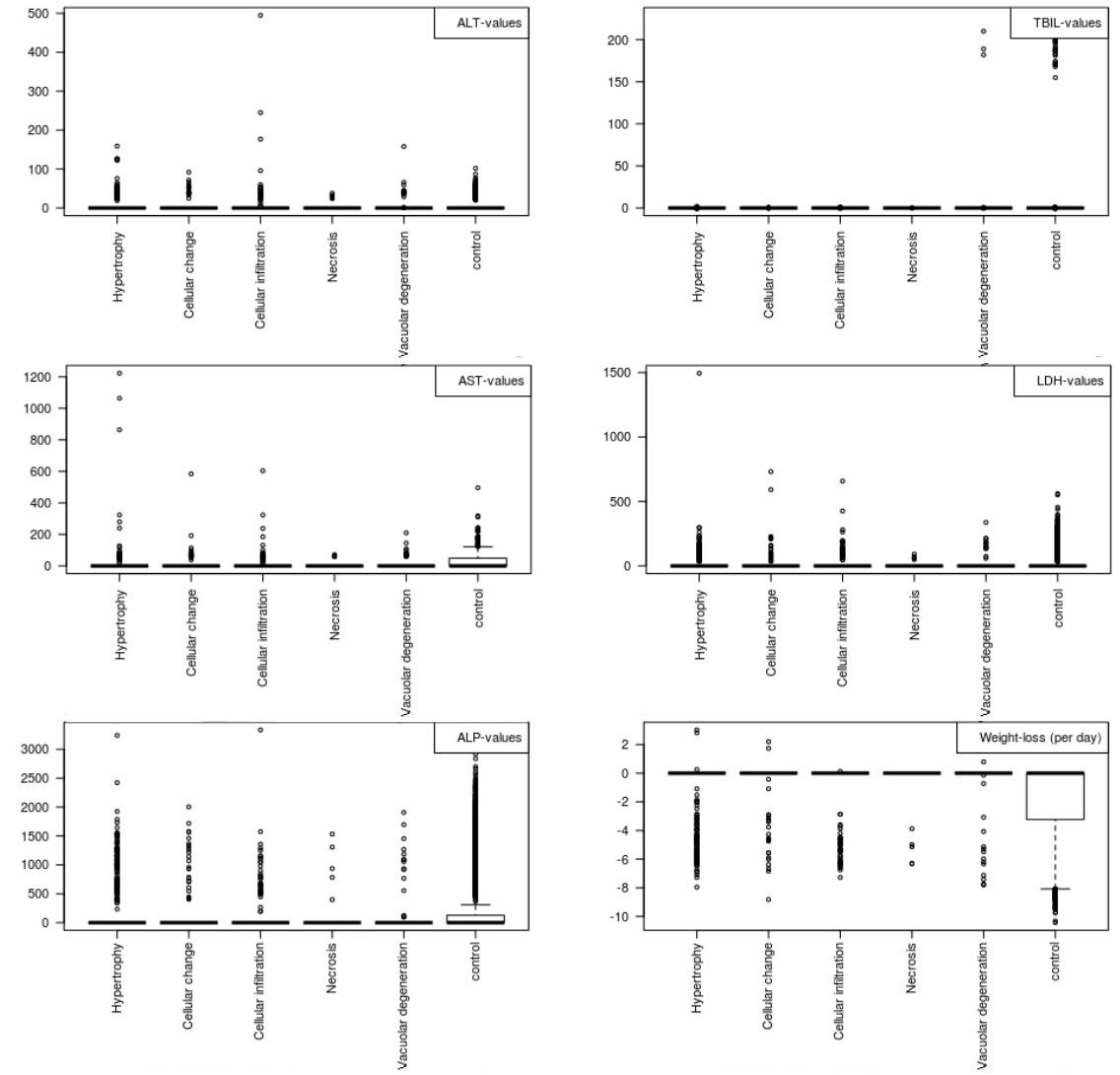


Figure 4.4: Boxplots showing deviation of gene expression values of those genes with highest variation in controls. (The bottom and top of the box are the 25th and 75th percentile, the band near the middle of the box is the 50th percentile, and the whiskers show the most extreme data points which are no more than 1.5 times the interquartile range from the box. Data points beyond this range (outliers) are marked with circles.



(a) (b)

Figure 4.5: Figures showing the correlation between histopathological finding and other phenotypical findings. Each point presents one array/rat. Phenotypical findings are (from top to bottom): (a) ALT, AST, ALP and (b) TBIL, LDH and weight loss per day. In each plot, the histopathological groups are (from left to right): Hypertrophy, Cellular change, Cellular infiltration, Necrosis, Vacuolar degeneration and controls.

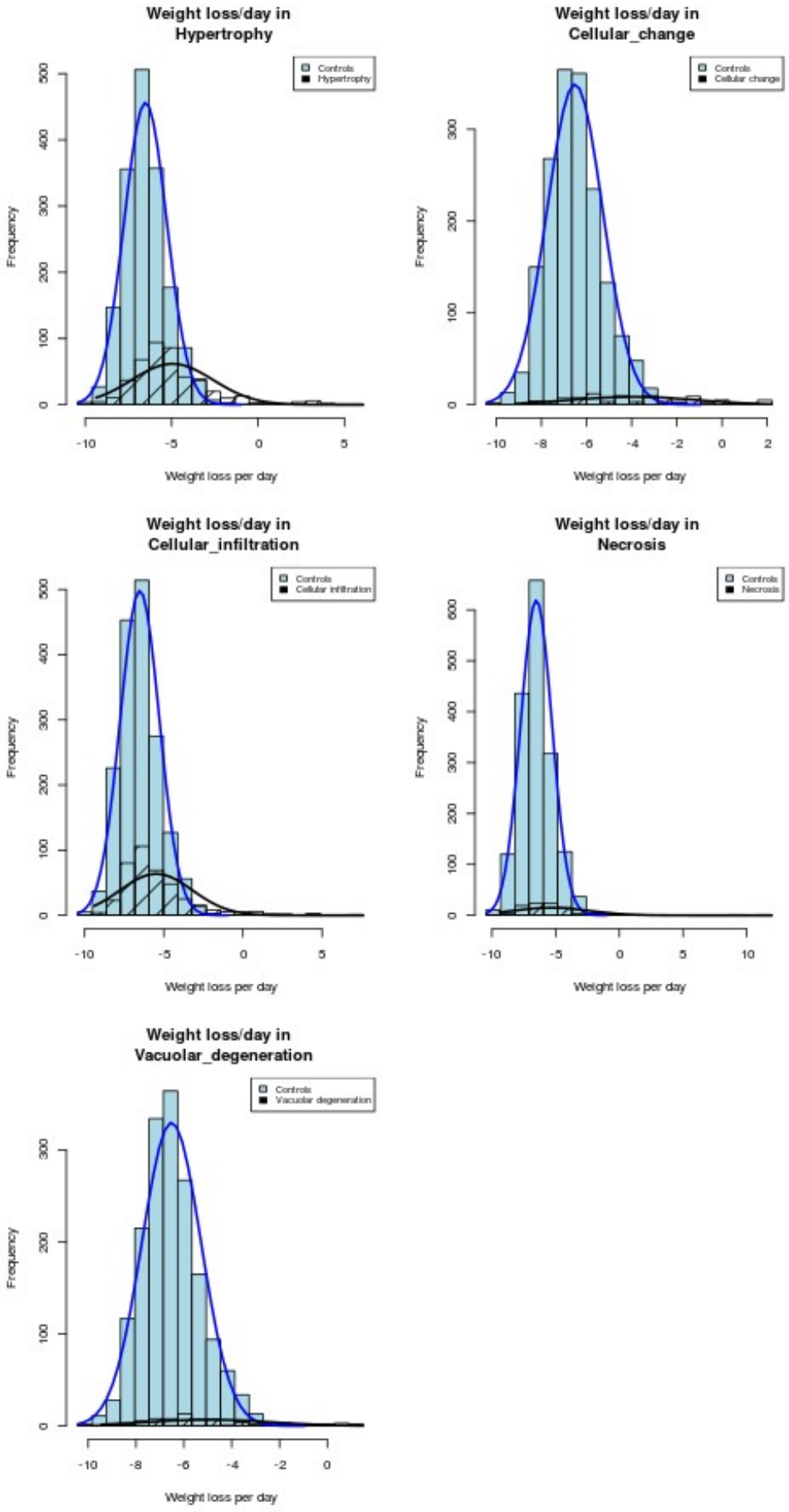


Figure 4.6: Weight loss of the rats with certain histopathological findings (black stripes) compared to the controls (blue). Normal curves are also shown to ease the comparison.

days with the lowest dose. On the other hand, some part of the samples seem to be getting further away from the controls.

To get a better idea of the factors causing the separation, samples were studied by separating dose levels and time points with different colors, as can be seen in Fig. 4.7 (upper row). It is difficult to say whether any correlation can be noticed, even when looking at the different time points in only a single (high) dose or different doses in one time point (14 days), see Fig. 4.7 (lower row). Analysis with other time points and dose levels alone produced somewhat similar results. Thus it can be determined that the type of the drug used has more effect on the separation than the dosage or the exposure time in general.

4.8 Dose Response

To validate the analysis and to see the effect of different doses to the gene expression, samples were organized as follows: first only those samples causing certain histopathological findings in more than five cases per compound when treated with highest dose were chosen. After this, samples treated with corresponding compounds and middle or low dose but which do not show any histopathological changes were chosen as their own groups. These three groups (high, middle and low) were analyzed against controls, and the computed p-values and (log) fold-changes were entered to Ingenuity. In Ingenuity IPA, a Comparison Analysis was performed between these groups.

In Fig. 4.8 one can see as an example the dose response comparison analysis of those samples causing vacuolar degeneration in the highest dose. Selected hepatotoxic functions are shown. The most obvious dose response behavior is seen in steatosis: highest dose shows lowest p-value (highest bar). This kind of behavior was expected, as those samples with fatty degeneration (which can be considered as a synonym for steatosis) were included in this histopathological group. Similar increasing trend can be seen in cholestasis, although the values are closer to the threshold.

On the other hand, the trend seems to be reversed with inflammation, glutathione depletion and damage. This seems controversial, as it would be expected to see more damage, oxidative stress and inflammation with higher doses (that is, to see more activation in genes responsible for these situations). There are few possible reasons behind this behavior. The dose might be that lethal to the cells that they are close to apoptosis or necrosis, and thus will not react to these situations. Expression of genes related to regeneration could give information about this situation, but in this case the p-values of these genes were under the threshold limit.

Another option is that cells or the whole organ may have adapted to the situation as well, and the change can hence be visualized in some other genes. Third and perhaps the most likely explanation would be that the levels of the proteins needed are already as high as possible or high enough considering the situation. Therefore the expression can return to normal level as new proteins are needed only to supplement the decayed ones. This situation reflects the challenges of toxicogenomics: gene expression values can

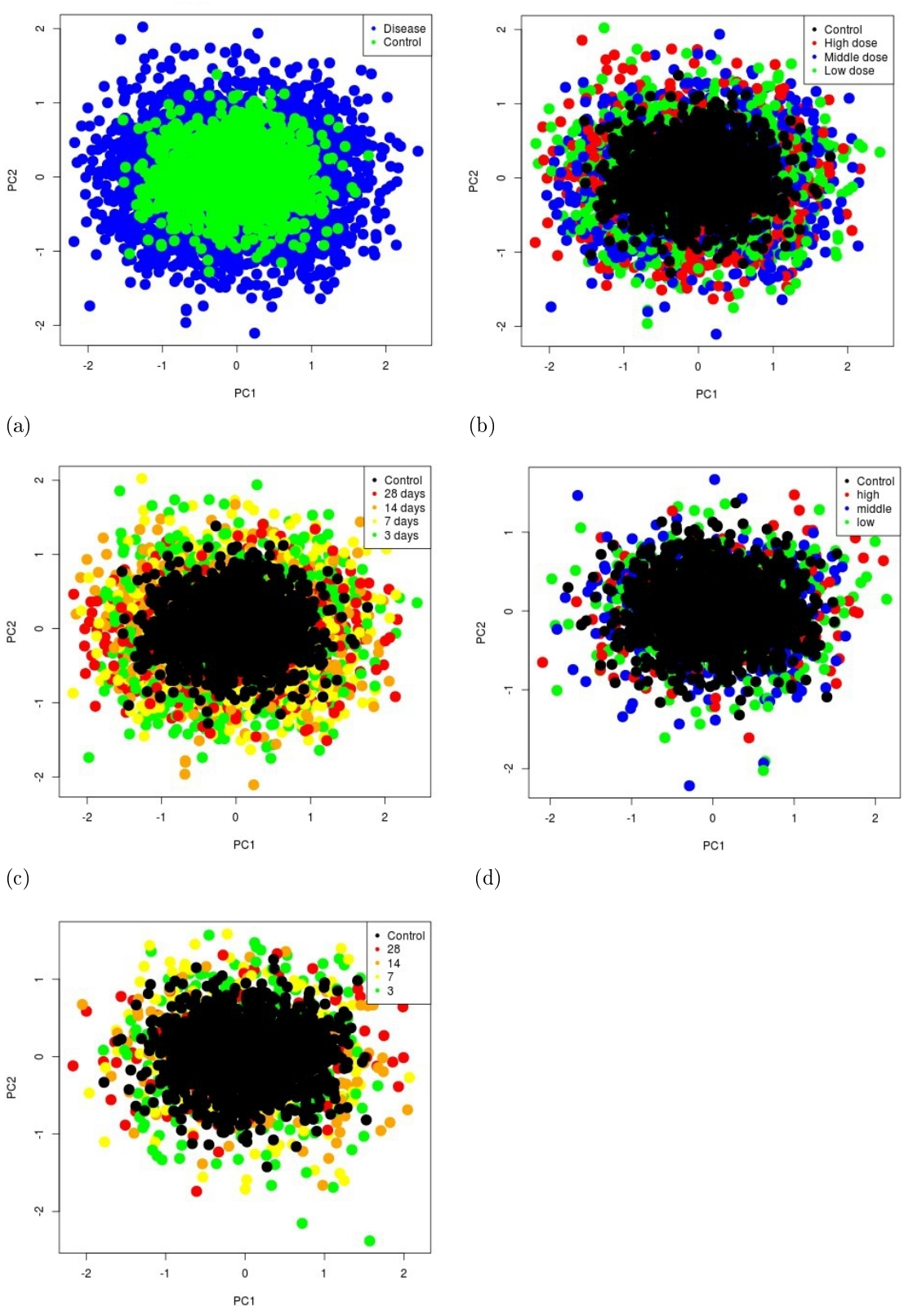


Figure 4.7: PCA plots showing (a) the separation between treated samples without any histopathological findings (blue or coloured) and controls (green or black), also with different dose levels (b) and exposure times (c) presented with different colors. Also those samples with 14 days exposure time (d) and those samples treated with the highest dose (e) are shown.

be somewhat miss-leading if one does not have any references or enough data.

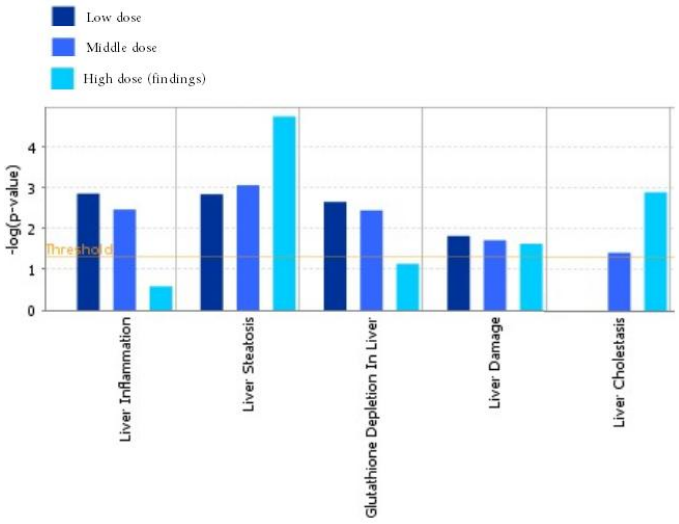


Figure 4.8: An example showing dose correlation (analysis done in Ingenuity IPA). Those samples showing vacuolar degeneration when treated with highest dose (lightest blue, bar on right) compared to those samples treated with same compounds (expected to cause vacuolar degeneration with higher doses or longer time) treated with middle and low dose (sky blue, in the middle and the darkest blue, on left). Notice that the scale in y-axis is $-log_{10}(p-value)$: the light orange threshold line is drawn at p-value = 0.05.

4.9 Time Response

Similarly to the dose response analysis described above, time response analysis with two time points (early = 3 and 7 days, late = 14 and 28 days) was prepared. Only the highest dose level was used, and the samples in late time point were required to show histopathological changes while the samples in the early time point were required to be clean. There was reasonable amount of data (i.e. enough samples) only in one group, hypertrophy.

Comparison Analysis was performed in Ingenuity IPA similarly as with dose response. The four time points of samples treated with aspirin were plotted in the same image for comparison. In Fig. 4.9 one can see some of the results (chosen Top Tox Functions).

Based solely on the two time points, it is difficult to say whether time response is actually seen or not. Nevertheless, compared to the dose response results, same kind of behavior is noticed in the same groups (for example steatosis is incresing as the dose or the exposure time is increased). However, the aspirin data seems to follow this assumed time behavior poorly: one must keep in mind that there are only three samples in each aspirin -group and thus more variation.

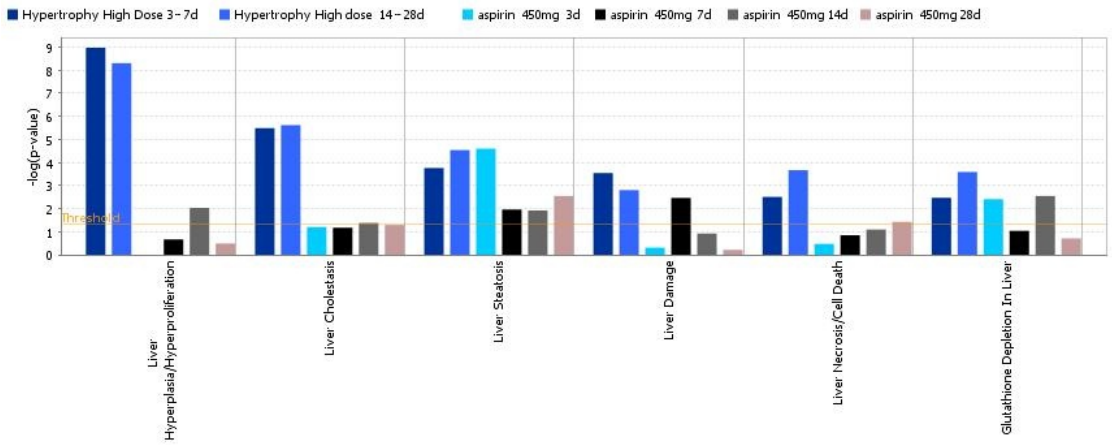


Figure 4.9: Comparison Analysis of two time points, early (3 and 7 days of treatment) and late (14 and 28 days) of samples showing hypertrophy in the later time point but not in the earlier. For comparison, all four time points treated with aspirin are also shown. Analysis done with Ingenuity IPA.

4.10 Use of the Data in Drug Development: Comparison to Few Examples

To see whether the data from the histopathology groups can be used to analyze the toxicity of possible drug candidates, a comparison analysis was performed in Ingenuity IPA using all of the five histopathology groups, two well-known drugs (ibuprofen and aspirin) and two drug candidates (referred here as molecules A and B). Some of the results from this analysis can be seen in Fig. 4.10.

Data of samples treated with ibuprofen and aspirin are from TG-GATEs. Data for molecules A and B was produced similarly. In all four cases, highest dose and longest available time-point was chosen in order to see maximal toxicity: for ibuprofen and aspirin the exposure time was 28 days and the doses were 200 mg/kg and 450 mg/kg, respectively. For molecules A and B the treatment time was 14 days and the doses were 600 mg/kg and 2000 mg/kg. One must keep in mind that these doses are ten times higher than the intended pharmacological dose for humans.

This data shows that the highest doses of even the well-known drugs can cause serious hepatotoxicity and damage to liver. Especially gene expression changes indicating steatosis is seen in all the four example samples treated with drug molecules. However, as we learned from the dose-response behavior in Chapter 4.8, this kind of presentation might be misleading, as we do not know whether some of the cases have ”gone past” of a certain stage. For example, it might be that the four drug molecule samples have already reached the maximum levels of those proteins that take care of glutathione depletion, and thus the bars are seen lower. On the other hand the situation might as well be that the cells in those samples are not suffering from oxidative stress, and therefore the corresponding gene expression levels are maintained relatively low. Hence the next step is to compare dose responses of the different groups and to see whether they show similar trends.

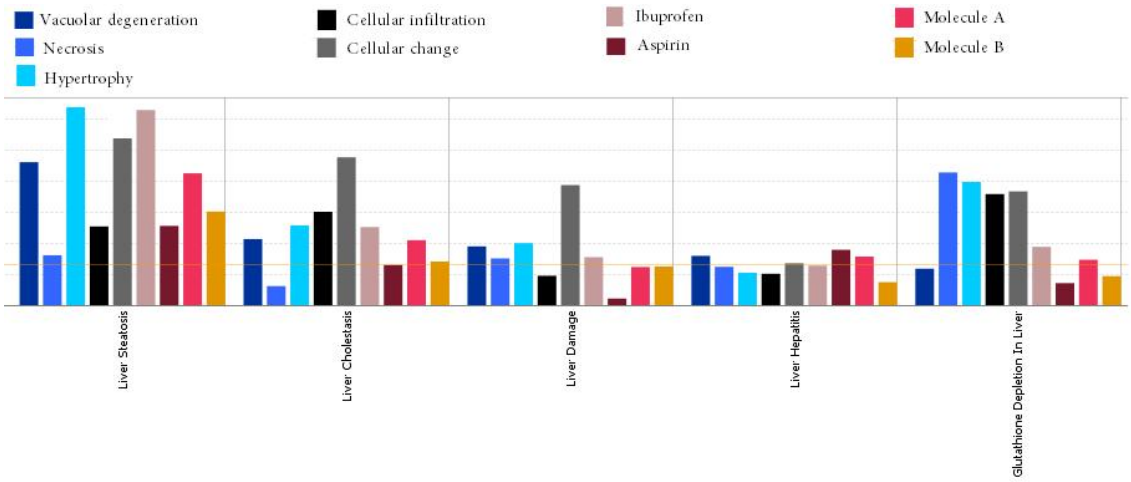


Figure 4.10: Here the data from different histopathology groups is compared to two well-known drugs (ibuprofen and aspirin) and two drug candidate molecules (referred here as molecule A and molecule B). Similarly to Fig. 4.8 some of the most interesting toxicity functions are shown. The scale in y-axis is $-log_{10}(p - value)$: the light orange threshold line is drawn at p-value = 0.05. Image produced with Ingenuity and modified.

In Fig. 4.11, samples showing vacuolar degeneration in histopathology and treated with the three different dose levels (low, middle, high, corresponding to 1, 3 and 10 times the medical dose) are compared to data from samples treated with ibuprofen in similar doses (20 mg/kg, 60 mg/kg and 200 mg/kg). (Note: ibuprofen was reported to cause only slight (4 samples) or moderate (2 samples) increased mitosis in TG-GATEs.) Similar trends were noticed, and this data confirms that ibuprofen is likely to cause steatosis and cholestasis in higher dose levels, as was first speculated based on the previous results (Fig. 4.10).

On the other hand, it seems that the case might be different with glutathione depletion: whereas in the samples treated with compounds causing vacuolar degeneration seem to defend against oxidative stress even with the lowest dose and not as much with the higher doses, the samples treated with ibuprofen seem to show increase in expression of corresponding genes only with the highest dose. The situation seems to be similar also with genes indicating liver damage, which would support this theory. This case is however more ambiguous as the p-values are closer to the threshold.

4.11 Comparison to Other Results

Same *in vivo* liver -data from TG-GATEs was used in a publication by Low et al ([39]), where they studied classification and prediction of hepatotoxicity using toxicogenomics and chemical features. They classified the data of 127 compounds in the database as hepatotoxic or nonhepatotoxic based on the histopathological findings and in some cases based on serum chemistry. They found 53 hepatotoxic compounds and 74 nonhepatotoxic ones. [39]

Based on the gene expression analysis performed to these groups, they discovered three pathways leading to long-term toxicity: these are Hnf4α-, Myc- and Eif2-centered networks.

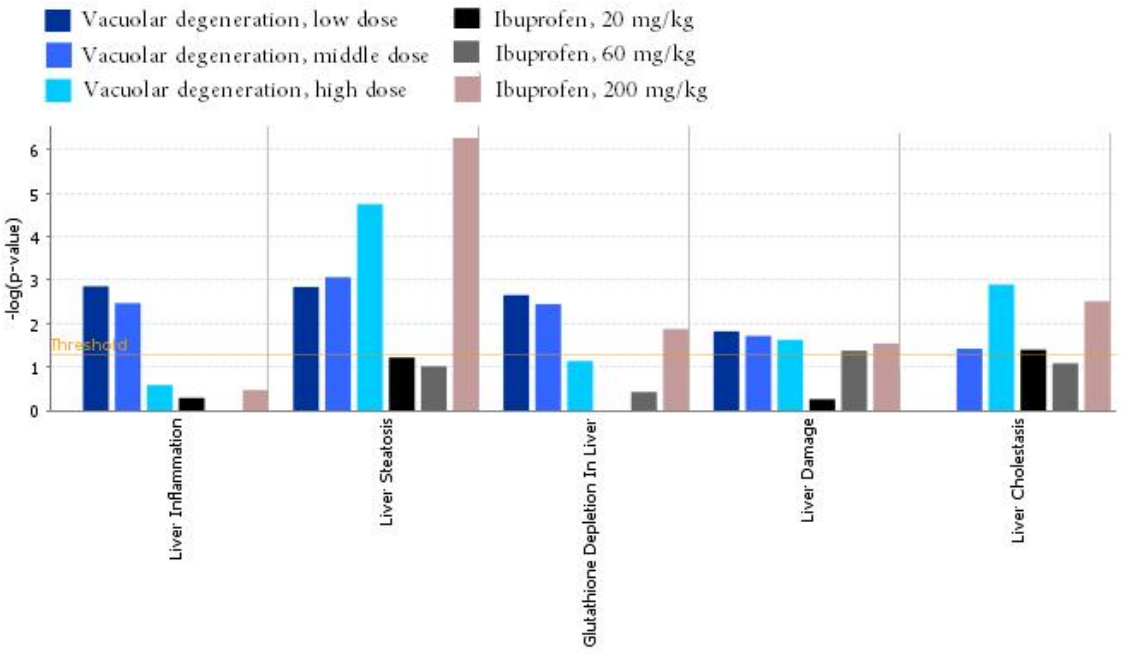


Figure 4.11: Comparison between one of the histopathology groups, vacuolar degeneration, and an example drug, ibuprofen, with the three dose levels. Ibuprofen samples were treated for 28 days. The scale in y-axis is $-log_{10}(p-value)$: the light orange threshold line is drawn at p-value = 0.05. Image produced with Ingenuity and modified.

Hnf4α is a transcriptional factor that functions in morphological and functional differentiation of hepatocytes, cell proliferation and detoxification, and it also has a function in the response to endoplasmic reticulum (ER) stress, which is caused by many toxicants. Also Eif2 signaling pathway seems to be central in liver damage through ER stress. Myc is transcriptional factor regulating cell proliferation, differentiation and apoptosis. In the study by Low et al [39] they show that both Hnf4α and Myc -centered pathways contain several up-regulated genes, which indicates that liver cells are feeling damage causing stress (see Fig. A.5 B borrowed from the publication in Appendixes). [39]

To compare our analysis with these results and to see whether some of the histopathological groups show particular up-regulation in the Myc-centered network, exactly the same network was build using Ingenuity IPA. After this, the data from the analysis of each histopathological group was overlayed on the network. Results and the original network from the study by Low et al [39] can be seen in the Appendixes, Fig. A.5.

Altogether the networks of the five histopathology groups show rather similar behavior, and many molecules in the Myc-network are, as expected, up-regulated. These include for example nucleophosmin 1 (Npm1) and TATA box binding protein -associated factor (Taf9), which are mentioned in the publication [39]. However, the general transcription factor IIIC polypeptide 3 (Gtf3c3), which was (according to the publication) expected to be up-regulated, showed no signs of specific regulation. Some components show also unexpected down-regulation, for example glucocorticoid nuclear receptor Nr3c1, which shows clear down-regulation in all groups. As the activation of Myc-centered network

indicates long-term toxicity and general stress, it was expected to see very little variation between the different hepatotoxic groups.

This comparison shows that this kind of networks indicating general hepatotoxicity can be used to support the classification between hepatotoxic and non-hepatotoxic, but regulation of single transcription factors (like Gtf3c3 or Nr3c1) can be misleading. Same kind of analysis was also performed for some samples treated with drugs that were known to be toxic in high doses (for example aspirin and WY-14643, which were classified toxic by Low et al [39]), and in some cases only some activation was noticed (aspirin), but some cases show high regulation even in the lowest doses and shortest exposures (WY-14643).

4.12 Compounds Responsible for Histopathological Findings

For the analysis purposes the compounds causing each histopathological finding were also studied (Fig. 4.12). It was noticed that some of the findings are caused by many different compounds (hypertrophy, cellular infiltration) while others are caused only by few (cellular change, vacuolar degeneration). In all groups some "noise" was noticed: there are several compounds noticed to cause one type of phenotype in only one or two samples. There were also several controls (tens in each case) with histopathological findings reported: these were excluded from the analysis.

These samples reflect the individual differences between the test organisms, and even though they may contain valuable information about different responses to the compounds, they might be problematic in analysis of the histopathology groups. As all the available data treated with a compound and reported to have histopathological findings in liver were used in the analysis, these "noise" samples are likely to cause variation in the results. This can be seen for example in PCA plots.

Some compounds stand out in more than one group (for example simvastatin, diazebam). In fact most of the samples reported to contain some phenotypical changes had several different histopathological findings: here only the most severe (or, in case of a tie, the first in alphabetical order) was chosen as the phenotype. This partly explains the similarity between the groups, but as can be seen from the Comparison Analyses (Chapter 4.2), the grouping used here also manages to emphasize the differences between the groups.

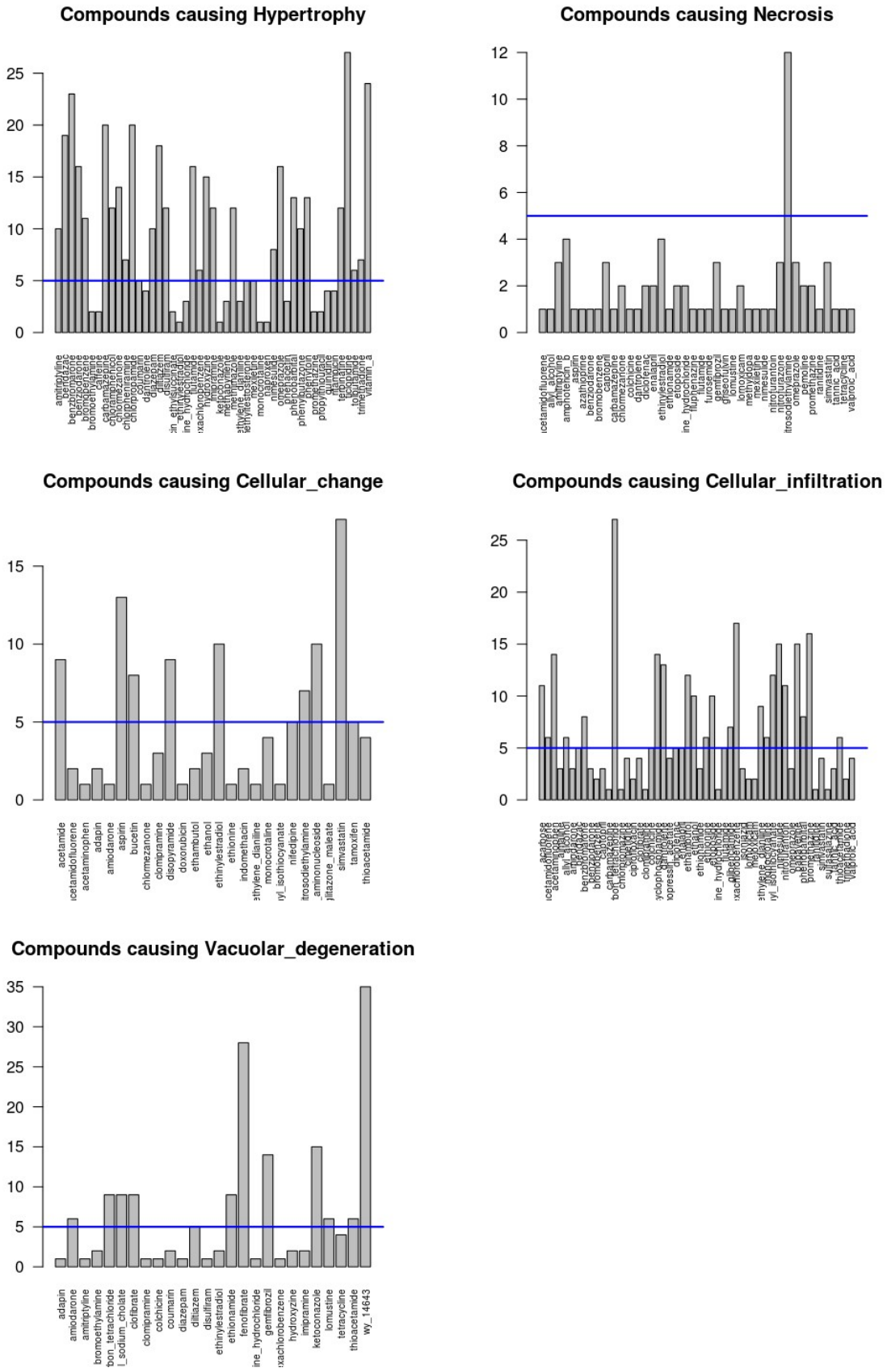


Figure 4.12: Different compounds causing certain histopathological findings when treated with any dose (low, middle or high).

5. DISCUSSION

In this chapter the results are briefly discussed and some conclusions are drawn based on them. Possible use of the knowledge gained in this study are speculated. Problems observed while analyzing the results are discussed, as well as possible future directions.

5.1 Summaries of the Results

Based on the PCA and KNN results (see Chapters 4.3 and 4.4) it was noticed that while some histopathology groups are very clearly separated from the controls (for example hypertrophy and vacuolar degeneration), others seem to be nearly impossible to detect (for example necrosis and cellular infiltration). One likely explanation for poor separation is that both with necrosis and cellular infiltration mostly minimal findings were reported (see Fig. A.4), and thus those samples were closer to controls. The deviation in necrosis-group is also explained by the fact that necrosis is caused by several different drugs with different mechanisms.

Another probable reason for poor separation and wide spread between samples with cellular infiltration and the controls is that the definition of the group is quite wide: it contains all reported cellular infiltrations, ground glass appearances and microgranulomas. The explanation might also be partly biological: gene expression might have a larger role in some of the histopathologies, whereas others might be visible in other regulatory systems. This is supported by the fact that the number of differently expressed genes does not correlate with the amount of data (number of samples) in that group.

Based on Fig. 4.12, the amount of different compounds causing the certain phenotype does not seem to correlate with the spread of the samples in PCA images: for example hypertrophy is caused by 47 different compounds, but the group is still very tight in the PCA image. This might indicate that the mechanisms causing hypertrophy are quite similar regardless of the compound causing it.

Even though vacuolar degeneration was as a group nicely separated from the controls (88 % sensitivity), PCA image seems to show several smaller clusters within the group. One explanation is the different, more detailed findings grouped together (see 5.1): clearly part of the samples reported to have eosinophilic granular degeneration seems to form a cluster of their own. This indicates that there might be two different mechanisms causing this finding.

Each of the five histopathological groups were studied by listing the differentially expressed genes and enriched GO gategories, and by performing Core and Comparison Analyses with Ingenuity IPA. The analysis done using Ingenuity IPA (see Fig. 4.1 and Core

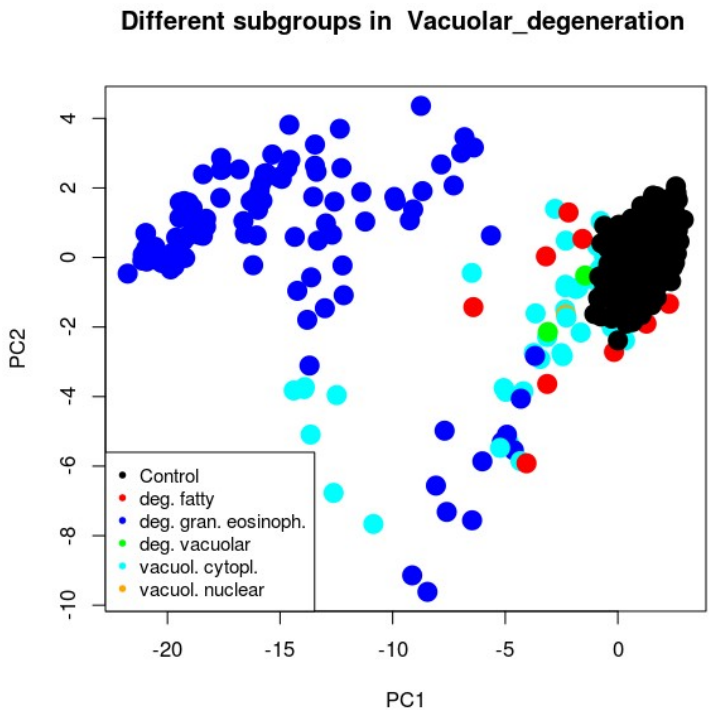


Figure 5.1: PCA image showing the more precise histopathological terms in vacuolar degeneration -group. (Deg. is short for degeneration, gran = granular, eosinoph. = eosinophilic, cytopl. = cytoplasmic and vacuol. = vacuolization/vacuolation)

Analysis -tables in Appendixes) reveals that especially the groups vacuolar degeneration, hypertrophy and cellular change seem to activate pathways related to steatosis. Besides this, cellular change -group seems to activate cholestasis, and hypertrophy -group glutathione depletion related pathways. These results are as expected. None of the group seem to highly activate liver damage or cell death -related pathways, which is assumed to be due to the long lasted situation: these genes have reached their expression peaks earlier.

Markers of general toxicity were also studied: gene expression of each of the five groups were overlayed on a Myc-centered pathway, which is assumed to be related to toxicity. It seems based on the result images (see Fig. A.5) that each group activates this pathway. This supports the use of this kind of indicator pathways for general toxicity.

No clear correlation between different histopathological findings and blood biochemistry values nor weight-loss of the rats were noticed. In human these blood values usually reflect some events taking place in the liver, but the case might be different with rats.

It was also noticed that the deviation of the gene expression values varies between genes. In some genes the expression values from 1392 control samples were really close to each other, whereas with other genes the spread was very large. Those samples with less variation in the controls are more likely to function as possible biomarkers. This kind of upper and lower limit -analysis would ease the estimation of severities of abnormal situations in the samples.

Based on the dose response analysis (see Chapter 4.8) performed to vacuolar degenera-

tion -group it can be concluded that in most cases the dose behavior seems evident. As it was possible to analyze time behavior with only two time points and using only one of the groups (Chapter 4.9), it is challenging to conclude whether some response is actually seen or not. Some dose and time behavior is also visible in the PCA images (Fig. A.2 and Fig. A.3), especially with hypertrophy -samples. As both dose and time response comparisons contained data with no phenotypical findings as the lower doses or earlier timepoint, these results encourage to use gene expression activity of certain pathways or gene groups as possible early biomerkers.

Some comparisons of certain well-known drugs and some possible drug candidates were performed against the five data groups. In Fig. 4.10 one can see that the data can be used as a benchmark when estimating possible hepatotoxicity of new (or old) drug molecules. As the data is quite large, the results can be considered quite reliable.

5.2 Relevance of the Results

As mentioned in Chapter 5.1, the analyzed data of the five histopathology groups can be used as benchmarks for the amount of changes in gene expression. By comparing new, similarly produced and analyzed microarray data to gene expression values of certain pathways and groups (for example those highlighted in this report), some conclusions can be drawn about possible hepatotoxicity of these compounds.

Especially the dose and time response data provide nice opportunities for comparison analyses. Based on these analyses, it seems that it might be possible to see some indicators of developing histopathological findings before they are seen in the phenotype (predictive use). Some time points or dose levels are needed from the new analyses as well to see the development of the situation: from single "snapshots" it is impossible to say whether the gene expression is at its highest peak, already sinking or just random fluctuation. Thus use of only toxicogenomics in decision making might be difficult and erroneous.

Ingenuity and GO enrichment analysis allows one to study the mechanisms of different hepatotoxicities in gene level. This could be useful when developing and modifying the drug to a less toxic direction.

Usually gene expression studies are challenged by the dimensions of the data: samples treated in certain way are compared to controls. The knowledge about the actual situation in the treated sample (its phenotype) might be very insufficient. Here we know the phenotype, and the transcript profiles could be explained based on this. This and the fact that the data is relatively large and the analyses are performed systematically (same arrays and protocols used etc.) makes these five groups a strong baseline to any hepatotoxicity analysis.

When performing the Ingenuity IPA analyses, it was noted that the results were in some cases slightly unexpected: for example, it had been expected to see higher activation in genes related to hypertrophy in hypertrophy-group. This situation might reflect the difference between a specific microarray study (using *in vivo* rats liver samples) and text-mining approach used when creating the Ingenuity-database. For example Ingenuity includes in

the analyses all available data from different organisms (human, rat, mouse etc.), different organs (liver, kidney, brain etc.) and different kind of studies (*in vivo*, *in vitro*). Thus the results must be used only as directional and the information about the treatment and phenotype of the sample are needed in the analysis. This also points out that it is wise to use several tools and databases instead of only trusting one.

5.3 Other Observations

Besides the results reported in Chapter 4, many general observations were made during the data analysis process. Here the complexity of genomics, use of rat as the main model organism in drug development are discussed. Some concerns and error sources are also presented, as well as future perspectives and other possible techniques.

5.3.1 Complexity of Genomics

Molecular networks and molecular/protein/gene interactions increase the complexity of the gene expression process, which adds challenges to toxicogenomics. By analyzing mRNA levels we get information of which genes are up- or downregulated, but we do not know which mRNAs are finally translated into proteins, how fast this happens, how many proteins are produced from one mRNA, how soon the mRNA is degraded, how much of a certain protein is needed to make a notable change to the situation in the cell and so on. Some regulation happens only on the protein level. Due to the networks, pathways and compensatory effects the direction of the regulation might not be that informative either. Besides this, one must remember that the mRNA is gathered from a sample that contains multiple cells: the values are averages over different cells at slightly different environments and situations, and perhaps of different cell types as well.

Thus, instead of actual values (compared to for example hemoglobin value in blood, with know boundaries and understandable unit), we get *probabilities* that *something* has happened. By taking advantage of the biological knowledge gathered in databases about networks, pathways and molecule interactions and their relations to different situations (diseases and symptoms), we can relate these single probabilities into larger entities and figure out what has happened.

5.3.2 Rat as Model Organism

In the data, tens of samples treated with vehicles only (the controls) were reported to have some histopathological findings in liver. These samples were excluded from the analysis, but they do reflect the effect of the vehicle diet [65] as well as the differences between individual rats: some rats are more sensitive to develop abnormalities than others, even though the experimental arrangement is such that each rat should be treated similarly. Differences in genetics caused by the environment (epigenetics) and other individual differences caused by the environment and history are still visible. These differences are likely to be in much higher level between individual humans, which makes it more difficult to

predict whether the drug will cause toxic responses in some individuals. [53]

Rat has been used as a model organism in drug development for a long time. Reasons for this are the fast breeding and relatively easy handling, but also the durability of the rat: it will stand really high doses of toxic compounds. Toxicogenomics uses overly high doses in order to "force" the toxic effects, and therefore the animal has to tolerate them, but does this skew the results?

It could be that nothing implying toxicity is found from the rats, but in other test animals even the low doses cause symptoms: drug development is often criticized for "developing drugs for rats". Due to its several good qualities as test organism and since regulations (by FDA) require certain toxicity tests done with rats, it is likely to be used as the main non-clinical test animal in future as well. For this purpose, it would be really useful to study the possibilities to link toxicogenomic results from rats with other organisms as well.

TG-GATEs has made an effort towards this direction, as similar information is gathered from rat *in vivo* and *in vitro* tests and from human cell *in vitro* tests. Some early results from first attempts to bridge between *in vivo* and *in vitro* and between rat and human are reported in [67]. For example expression changes in genes related to cell proliferation and apoptosis were detected *in vivo* but not *in vitro*. Between rats and humans, the patterns of changes were similar, but the extent of the changes was greater in rat cells than in human cells.

5.3.3 Possible Sources of Errors and Other Concerns

Besides the aforementioned individual differences between the rats and their special durability, one should keep in mind that the grouping used in this project was based on visual inspection of the histopathological samples: two histopathologists examined the liver samples with a microscope and graded the findings according to certain instructions. Even though this was done systematically and in blinded manner, the effect of chance (for example from which part the sample was taken, which areas were analyzed with the microscope) and possible lack of repeatability (due to individual opinions) can not be ignored.

Some universal standards for the histopathology and other terms could maybe improve these studies as well. Especially with histopathology terms some universal scoring system would also be helpful. Besides this, samples were chosen to belong to only one group even if several different findings were reported. However, based on the results it seems that the used grouping managed to highlight the differences between the groups.

5.4 Further Analyses and Future Perspectives

Next step in the analysis would be to further validate the use of the data as benchmark by testing the classification with an new sample with known histopathological findings. PCA, KNN classification and hierarchical clustering would show how this new sample settles in the data. By comparing the gene expression values of the new sample to the gene expression deviations of the controls and different groups would be interesting as well.

The new data should preferably be from a similar experiment, i.e. using Affymetrix Genechip (Rat Genome 230 2.0 Array) and rat liver *in vivo* -samples treated with some compound for 3-28 days. Controls and batch correction for this new data are also needed. It might be difficult to find data that fits in to this description, and thus some bridging between different microarrays and experiments would be really useful as well, especially as the techniques are constantly developing.

As stated before, the histopathology groups used in this study were quite broad: use of smaller groups (sub-groups) might bring more insight to the analysis. Another option would be to try to combine those drugs with similar mechanisms, target molecules or structure and analyze them as a group.

It would be interesting also to choose some of the differentially expressed genes and test their use as biomarkers. These genes could be chosen based on the gene expression value deviation plots shown in section 4.5 or based on the analyses performed with Ingenuity. Their reliability could be tested by using them in the classification and clustering. They could also be studied further using different databases and text-mining approach.

As stated before, some bridging is needed between the *in vivo* and *in vitro* experiments and between rat and human samples. TG-GATEs database is aiming towards this direction [67]: as there already is some rat and human *in vitro* data available from experiments performed with the same compounds and techniques, it would be interesting to compare them to the data used in this project.

One of the goals in toxicogenomics is to enable toxicity analyses with lower doses and shorter exposure times. This way smaller amount of the possible drug candidate would be sufficient, and time could be saved in synthesis as well as with the experiments. The results showed that based on the transcript profiles it might be possible to predict hepatotoxicity before it is seen in the phenotype. FDA and MAQC seem to support this kind of development.

Predictive use in toxicogenomics is not the only use of genomics in drug development. As stated before, transcript profiling can also be used to study the mechanisms of toxicity and function of the drug candidate. With help of this information, the drug molecule could be improved. Genomics could also be used in attempts to personalize medicine: certain single-nucleotide polymorphisms (SNPs) are known to be associated with certain kind of drug responses. [42, 53]

As stated by D.L.Mendrick in [43], biomarkers of toxicity are needed to both A) provide additional tools for the clinical management of patients and to enable preventive measures, and B) identify unsafe drugs earlier in development and predict the possible toxicity in many patients or certain individuals. However, tissue samples are rarely available, and hence it would be really useful to search biomarkers from blood samples. [43]

Even though peripheral blood cells can be used to study tissue damage or dysfunction due to diseases, there is rather little evidence of this kind of detection of toxicity. One example is the prediction of exposure to harmful levels of acetoaminophen from the expression signatures in rat [7]: these rat genes were translated to corresponding human genes (=orthologs) and successfully used to classify acetaminophen-toxicated patients from

controls. This is a very encouraging result suggesting that it would be wise to study the blood cell samples as well. [43]

5.4.1 Other Techniques

As discussed before, toxicogenomics and microarrays are mainly qualitative methods, and although they are widely used in different fields including toxicogenomics, their use has certain challenges. Several other methods that could be used in similar studies exist and some of them are presented here. In order to achieve confident research results, combinations of these methods should be used. These combinations and comparisons could also bring light to the knowledge of both techniques.

Proteomics, i.e. the "high throughput separation, display and identification of proteins", might prove as a at least equally useful technique for drug development as genomics which is based on transcripts, and it might not indicate much about the actual protein levels. Proteomics includes techniques from gel electrophoresis to mass spectrometry. The advantage of proteomics is that body fluids can be studied without need for cellular material, and since many proteins are secreted to the fluids, it could be possible to predict physiological states based on their amounts. [34]

Same advantages concern **metabolomics**. Metabolomics measures all or a subset of concentrations of small molecular weight metabolites in a certain system (for example body fluid) [33]. Usually a separation step such as chromatography or electrophoresis is performed first and then a identification step like mass spectrometry follows [33]. Especially with clinical environment protein or metabolite biomarker would be better, as they could be tested with immonoassays which are generally faster than genomic/genetic tests. On the other hand metabolomics is more sensitive to unwanted variation caused by different diets than gemomics. [43]

Genomics is a qualitative method: based on the results it can be said that something happened, but it is more difficult to say how serious the event was. Quantitative reverse transcriptase polymerase chain reaction (**QC RT-PCR**) would give a sense of magnitudes of the changes. In QR RT-PCR the mRNA samples are reversibly transcriped to DNA, which is then amplified. Using internal reference genes the exact amount of mRNA molecules can be determined. [23]

Transcriptome is the complete set of transcripts (including mRNAs, non-coding RNAs and small RNAs) in a cell and their quantities. To analyze the transcriptome, a method called RNA-Seq can be used. It is often referred to as next generation sequencing method. In RNA-Seq, a population of (fractioned) RNA is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing). After this the resulting reads are either aligned to reference transcripts, or a new genome scale transcription map is assembled *de novo*. [71, 44]

RNA-Seq is still under development, although several commercial systems have been published (Illumina, Applied Biosystems SOLiD, Roche 454 Life Science). It is likely to

be the next highly used technology in the field due to its many benefits: it is not limited by existing genomic sequences, it has very low background signal and high dynamic range (even greater than 9,000-fold range, compared to maximum of few-hundredfold range in microarrays) and it is highly accurate for quantifying expression levels. Challenges are related to bias caused by fragmenting, and to analysis and complexity of the data. [71]

6. CONCLUSIONS

The aim of this project was to study whether it is possible to recognize typical gene expression profiles for certain phenotypical (i.e. histopathological) changes in liver samples. Another goal was to study whether this data can be used as a reference when predicting possible histopathologies in future samples treated with possible drug candidates, and thus to aid in the search of suitable biomarkers of toxicity. These biomarkers would support early stage drug development. For this purpose, data from *in vivo* gene expression studies of rat liver from Japanese Toxicogenomics Project database (TG-GATEs) was used. Possibility of using smaller doses and shorter experiment times in order to save resources was studied by comparing different exposure times and dose levels.

Identification of histopathology groups was studied with PCA, KNN cross-validation and comparison of differentially expressed genes and enriched gene ontology classes and pathways. Some histopathology groups showed good separation from other groups and controls, whereas others were more spread. Spreading was likely to be caused by experimental factors, such as different dose levels, exposure times, severities of the findings and by the looseness in the definition of the group, but also different mechanisms causing a certain type of histopathology. As expected, genes in pathways related to steatosis, cholestasis and glutathione depletion were differentially expressed in most of the groups. Also pathways known to indicate toxicity (such as Myc-centered pathway) were activated.

From the comparisons between the histopathology groups and drug molecules (both well-known and possible drug candidates) it was noticed that the data can be used as a benchmark describing the levels of gene expression in hepatotoxicity. Similarly the deviation of the gene values in control group and in the histopathology groups can be used as a measure of changes in gene expression. The extend and systematic protocol of the TG-GATEs makes this data a strong baseline to analysis.

Treated samples without any histopathological findings were used in dose and time response analysis and some responsive behavior was noticed. This encourages the use of gene expression profiles as early biomarkers, as predictive changes were noticed in gene expression before anything was noticed in phenotype. However these studies also suggest that few time and dose points are needed in future analyses as well, as the direction of the changes is difficult to see from single "snapshots" of gene expression. No correlation between histopathological findings and blood biochemistry values were noticed, which reflects the difference between rat and human. Some correlation to weight-loss was noticed, which proves that the compounds were indeed harmful for the test animals.

When considering the results one should keep in mind that the samples may have contained different kinds of liver cells, and that the determination of the histopathological

groups was based on individual visual inspection. It is also advisable to keep in mind the complexity of gene expression and thus not rely heavily on single gene values or directions of the regulation, but rather to examine overall pictures gained by gene ontology enrichments or pathway analyses. The data can be used in estimation of toxicity as one measure amongst others and as a benchmark: it is relatively easy to compare future analysis results to the data using Ingenuity IPA. The results can also be used when studying mechanisms of toxicity, and hence possibly to guide the direction of development of a candidate drug molecule.

In the future, it would be interesting to study in a similar way the bridging between this *in vivo* data and corresponding *in vitro* data from TG-GATEs. Then also the differences and similarities between rat and human samples could be studied. Other histopathological groups could be studied as well, and the groups could be determined more specifically in order to get more precise information about each case. The use of the data as benchmark could also be further validated and tested by using it in classification of new samples with known histopathological findings. Some of the differentially expressed genes found in this project and the possibility of using them as biomarkers of hepatotoxicity could also be further studied. It might also be interesting to see the correlation of the findings presented here to gene expression measured from blood samples, as some encouraging results of this kind of studies exist.

REFERENCES

[1] Affymetrix. GeneChip Expression Analysis, Data Analysis Fundamentals (Manual). *Available at: http://media.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf*, 2012.

[2] Affymetrix. The genechip rat genome 230 arrays -datasheet. http://www.osa.sunysb.edu/udmf/ArraySheets/rat230_2_datasheet.pdf, June 2012.

[3] Affymetrix Inc. Statistical Algorithms Description Document. *Available at: http://media.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf*, 2002.

[4] Magnus Astrand. Contrast normalization of oligonucleotide arrays. *Journal of computational biology: Journal of computational molecular cell biology*, 10(1):95–102, January 2003.

[5] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing . *Journal of the Royal Statistical Society (Series B, Methodological)*, 57:289–300, 1995.

[6] B M Bolstad, R a Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, 19(2):185–93, January 2003.

[7] P R Bushel, A N Heinloth, J Li, L Huang, J W Chou, G A Boorman, D E Malarkey, C D Houle, S M Ward, R E Wilson, R D Fannin, M W Russo, P B Watkins, R W Tennant, and R S Paules. Blood gene expression signatures predict exposure levels. *PNAS*, 104(46):1–6, 2007.

[8] Seth Carbon, Amelia Ireland, Christopher J Mungall, ShengQiang Shu, Brad Marshall, and Suzanna Lewis. AmiGO: online access to ontology and annotation data. *Bioinformatics (Oxford, England)*, 25(2):288–9, January 2009.

[9] Marc Carlson, Seth Falcon, Herve Pages, and Nianhua Li. Genome wide annotation for rat (org.rn.eg.db), documentation. http://bioconductor.org/packages/2.6/data/annotation/html/org.Rn.eg.db.html, July 2012.

[10] Fred Hutchinson Cancer Research Center. Bioconducor, Open source software for Bioinformatics. http://www.bioconductor.org/, August 2012.

[11] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLoS ONE*, 6:1–10, 2011.

[12] Manhong Dai, Pinglang Wang, Andrew D Boyd, Georgi Kostov, Brian Athey, Edward G Jones, William E Bunney, Richard M Myers, Terry P Speed, Huda Akil, Stanley J Watson, and Fan Meng. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic acids research*, 33(20):e175, January 2005.

[13] Affymetrix developer network. Affymetrix .CEL -Data File Format. http://www.stat.lsa.umich.edu/~kshedden/Courses/Stat545/Notes/AffxFileFormats/cel.html, June 2012.

[14] S. Dudoit, R. C. Gentleman, and J. Quackenbush. Open Source Software for the Analysis of Microarray Data. *BioTechniques*, 34:S45–S51, 2003.

[15] Editorial. Making the most of microarrays. *Nature Biotechnology*, 24(9):1039, 2006.

[16] The R Project for Statistical Computing. Project webpage. http://www.r-project.org/, June 2012.

[17] Genomics Technology Frontiers In Genetics, Services. Technology and equipment -page, affymetrix genechip system. http://www.frontiers-in-genetics.org/page.php?id=gen-technology_en, August 2012.

[18] Laurent Gautier, Rafael Irizarry, Leslie Cope, and Ben Bolstad. Description of affy. *Bioconductor*.

[19] R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Duboit. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, 2005.

[20] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, January 2004.

[21] S. Gentleman, R., Falcon. Tools for manipulating GO and microarrays: GOstats. *Available at: http://www.bioconductor.org/packages/release/bioc/html/GOstats.html*.

[22] Edoardo G Giannini, Roberto Testa, and Vincenzo Savarino. Liver enzyme alteration: a guide for clinicians. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 172(3):367–79, February 2005.

[23] Ursula Gibson, Christian Heid, and Mickey Williams. A novel method for real time quantitative RT-PCR. *Genome Research*, 6(10):995–1001, October 1996.

[24] F. Hahne, W. Huber, R. Gentleman, and S. Falcon. *Bioconductor Case Studies*. Springer, 2008.

[25] M. J. Heller. DNA Microarray Technology: Devices, Systems and Applications. (Review). *Annu. Rev. Biomed. Eng.*, 4:129–153, 2002.

[26] Michael J Heller. DNA microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4:129–53, January 2002.

[27] I. Hovatta, K. Kimppa, A. Lehmussola, T. Pasanen, J. Saarela, I. Saarikko, J. Saharinen, P. Tiikkanen, T. Toivanen, M. Tolvanen, M. Vihinen, G. Wong, J. Tuimala, and M. Laine. *DNA Microarray Data Analysis.* CSC -Scientific Computing Ltd., 2005.

[28] Earl Hubbell, Wei-min Liu, and Rui Mei. Robust estimators for expression analysis. 18(12):1585–1592, 2002.

[29] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31:8, 2003.

[30] Rafael a Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2):249–64, April 2003.

[31] Ian B Jeffery, Desmond G Higgins, and Aedín C Culhane. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC bioinformatics*, 7:359, January 2006.

[32] P.R. Wheater H.G.Burkitt A.Stevens J.S.Lowe. *Basic Histopathology.* Churchill Livingstone, 1985.

[33] Douglas B Kell. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug discovery today*, 11(23-24):1085–92, December 2006.

[34] Sandy Kennedy. The Role of Proteomics in Toxicology : Identification of Biomarkers of Toxicity by Protein Expression Analysis. *Biomarkers*, 7(4):269–290, 2002.

[35] N. Kiyosawa, Y. Ando, K. Watanabe, N. Niino, S. Manabe, and T. Yamoto. Scoring Multiple Toxicological Endpoints Using a Toxicogenomic Database. *Toxicology Letters*, 188:91–97, 2009.

[36] William M. Lee. Drug-induced hepatotoxicity. *New England Journal of Medicine*, 349(5):474–485, 2003.

[37] Y. F. Leung and D. Cavalieri. Fundamentals of cDNA microarray data analysis (Review). *Trends in Genetics*, 19:649–659, 2003.

[38] C. Li and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. . *Oxford Journals: Life Sciences & Mathematics & Physical Sciences: Biostatistics.*, 8:118–127, 2006.

[39] Y. Low, T. Uehara, Y. Minowa, H. Yamada, Y. Ohno, T. Urushidani, A. Sedykh, E. Muratov, V. Kuz'min, D. Fourches, H. Zhu, I. Rusyn, and A. Tropsha. Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. *Chemical Research In Toxicology*, 24:1251–1262, 2011.

[40] J. Luo, M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, T. Shi, W. Tong, L. Shi, H. Hong, C. Zha, F. Elloumi, W. Shi, R. Thomas, S. Lin, G. Tillinghast, G. Liu, Y. Zhou, D. Herman, Y. Li, Y. Deng, H. Fang, P. Bushel, M. Woods, and J. Zhang. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. . *The Pharmacogenomics Journal*, 10:278–291, 2010.

[41] Emily Mankin. Principal Components Analysis: A How-To Manual for R. https://parmesan.mc.vanderbilt.edu/turnersd/ggd/2010-11-18-PCA-tutorial.pdf, June 2012.

[42] J J McCarthy and R Hilfiker. The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nature biotechnology*, 18(5):505–8, May 2000.

[43] Donna L Mendrick. Genomic and genetic biomarkers of toxicity. *Toxicology*, 245(3):175–81, March 2008.

[44] Michael L Metzker. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46, January 2010.

[45] Molecular and University of Michigan Behavioral Neuroscience Institute, Microarray Lab. BRAINARRAY: Custom CDF files. http://brainarray.mbni.med.umich.edu/brainarray/Database/CustomCDF /genomic_curated_CDF.asp, June 2012.

[46] D. L. Nelson and M. M. Cox. *Lehninger: Principles of Biochemistry*. W.H. Freeman and Company, 2008.

[47] National Institute of Biomedical Innovation (NIBIO). TG-GATEs database. http://toxico.nibio.go.jp/open-tggates/search.html, August 2012.

[48] The Gene Ontology. AmiGO webpage. http://amigo.geneontology.org/cgi-bin/amigo/go.cgi, June 2012.

[49] Stuart D Pepper, Emma K Saunders, Laura E Edwards, Claire L Wilson, and Crispin J Miller. algorithms. 12:1–12, 2007.

[50] J Quackenbush. Computational analysis of microarray data. *Nature reviews. Genetics*, 2(6):418–27, June 2001.

[51] Benjamin Milo Bolstad Rafael A. Irizarry, Laurent Gautier and Crispin Miller. Affy, Methods for Affymetrix Oligonucleotide Arrays. *Available at: http://www.bioconductor.org/packages/release/bioc/html/affy.html.*

[52] Brian Ripley. R Documentation: Package 'class ' (version 7.3-4). Functions for classification. *Available at: http://cran.r-project.org/web/packages/class/class.pdf.*

[53] Allen D. Roses. Pharmacogenetics and the Practice of Medicine. *Nature*, 405(6788):857–65, June 2000.

[54] E. E. Schadt, C. Li, C. Su, and W. H. Wong. Analyzing High-Density Oligonucleotide Gene Expression Array Data. *Journal of Cellular Biochemistry*, 80:192–202, 2000.

[55] Leming Shi, Gregory Campbell, Wendell D Jones, Fabien Campagne, Zhining Wen, Stephen J Walker, Zhenqiang Su, Tzu-Ming Chu, Federico M Goodsaid, Lajos Pusztai, John D Shaughnessy, André Oberthuer, Russell S Thomas, Richard S Paules, Mark Fielden, Bart Barlogie, Weijie Chen, Pan Du, Matthias Fischer, Cesare Furlanello, Brandon D Gallas, Xijin Ge, Dalila B Megherbi, W Fraser Symmans, May D Wang, John Zhang, Hans Bitter, Benedikt Brors, Pierre R Bushel, Max Bylesjo, Minjun Chen, Jie Cheng, Jing Cheng, Jeff Chou, Timothy S Davison, Mauro Delorenzi, Youping Deng, Viswanath Devanarayan, David J Dix, Joaquin Dopazo, Kevin C Dorff, Fathi Elloumi, Jianqing Fan, Shicai Fan, Xiaohui Fan, Hong Fang, Nina Gonzaludo, Kenneth R Hess, Huixiao Hong, Jun Huan, Rafael a Irizarry, Richard Judson, Dilafruz Juraeva, Samir Lababidi, Christophe G Lambert, Li Li, Yanen Li, Zhen Li, Simon M Lin, Guozhen Liu, Edward K Lobenhofer, Jun Luo, Wen Luo, Matthew N McCall, Yuri Nikolsky, Gene a Pennello, Roger G Perkins, Reena Philip, Vlad Popovici, Nathan D Price, Feng Qian, Andreas Scherer, Tieliu Shi, Weiwei Shi, Jaeyun Sung, Danielle Thierry-Mieg, Jean Thierry-Mieg, Venkata Thodima, Johan Trygg, Lakshmi Vishnuvajjala, Sue Jane Wang, Jianping Wu, Yichao Wu, Qian Xie, Waleed a Yousef, Liang Zhang, Xuegong Zhang, Sheng Zhong, Yiming Zhou, Sheng Zhu, Dhivya Arasappan, Wenjun Bao, Anne Bergstrom Lucas, Frank Berthold, Richard J Brennan, Andreas Buness, Jennifer G Catalano, Chang Chang, Rong Chen, Yiyu Cheng, Jian Cui, Wendy Czika, Francesca Demichelis, Xutao Deng, Damir Dosymbekov, Roland Eils, Yang Feng, Jennifer Fostel, Stephanie Fulmer-Smentek, James C Fuscoe, Laurent Gatto, Weigong Ge, Darlene R Goldstein, Li Guo, Donald N Halbert, Jing Han, Stephen C Harris, Christos Hatzis, Damir Herman, Jianping Huang, Roderick V Jensen, Rui Jiang, Charles D Johnson, Giuseppe Jurman, Yvonne Kahlert, Sadik a Khuder, Matthias Kohl, Jianying Li, Menglong Li, Quan-Zhen Li, Shao Li, Zhiguang Li, Jie Liu, Ying Liu, Zhichao Liu, Lu Meng, Manuel Madera, Francisco Martinez-Murillo, Ignacio Medina, Joseph Meehan, Kelci Miclaus, Richard a Moffitt, David Montaner, Piali Mukherjee, George J Mulligan, Padraic Neville, Tatiana Nikolskaya, Baitang Ning, Grier P Page, Joel Parker, R Mitchell Parry, Xuejun Peng, Ron L Peterson, John H Phan, Brian Quanz, Yi Ren, Samantha Riccadonna, Alan H Roter, Frank W Samuelson, Martin M Schumacher, Joseph D Shambaugh, Qiang Shi, Richard Shippy, Shengzhu Si, Aaron Smalter, Christos Sotiriou, Mat Soukup, Frank Staedtler, Guido Steiner, Todd H Stokes, Qinglan Sun, Pei-Yi Tan, Rong Tang, Zivana Tezak, Brett Thorn, Marina Tsyganova, Yaron Turpaz, Silvia C Vega, Roberto

Visintainer, Juergen von Frese, Charles Wang, Eric Wang, Junwei Wang, Wei Wang, Frank Westermann, James C Willey, Matthew Woods, Shujian Wu, Nianqing Xiao, Joshua Xu, Lei Xu, Lun Yang, Xiao Zeng, Jialu Zhang, Li Zhang, Min Zhang, Chen Zhao, Raj K Puri, Uwe Scherf, Weida Tong, and Russell D Wolfinger. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, 28(8):827–38, August 2010.

[56] Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet a Warrington, Shawn C Baker, Patrick J Collins, Francoise de Longueville, Ernest S Kawasaki, Kathleen Y Lee, Yuling Luo, Yongming Andrew Sun, James C Willey, Robert a Setterquist, Gavin M Fischer, Weida Tong, Yvonne P Dragan, David J Dix, Felix W Frueh, Frederico M Goodsaid, Damir Herman, Roderick V Jensen, Charles D Johnson, Edward K Lobenhofer, Raj K Puri, Uwe Schrf, Jean Thierry-Mieg, Charles Wang, Mike Wilson, Paul K Wolber, Lu Zhang, Shashi Amur, Wenjun Bao, Catalin C Barbacioru, Anne Bergstrom Lucas, Vincent Bertholet, Cecilie Boysen, Bud Bromley, Donna Brown, Alan Brunner, Roger Canales, Xiaoxi Megan Cao, Thomas a Cebula, James J Chen, Jing Cheng, Tzu-Ming Chu, Eugene Chudin, John Corson, J Christopher Corton, Lisa J Croner, Christopher Davies, Timothy S Davison, Glenda Delenstarr, Xutao Deng, David Dorris, Aron C Eklund, Xiao-hui Fan, Hong Fang, Stephanie Fulmer-Smentek, James C Fuscoe, Kathryn Gallagher, Weigong Ge, Lei Guo, Xu Guo, Janet Hager, Paul K Haje, Jing Han, Tao Han, Heather C Harbottle, Stephen C Harris, Eli Hatchwell, Craig a Hauser, Susan Hester, Huixiao Hong, Patrick Hurban, Scott a Jackson, Hanlee Ji, Charles R Knight, Winston P Kuo, J Eugene LeClerc, Shawn Levy, Quan-Zhen Li, Chunmei Liu, Ying Liu, Michael J Lombardi, Yunqing Ma, Scott R Magnuson, Botoul Maqsodi, Tim McDaniel, Nan Mei, Ola Myklebost, Baitang Ning, Natalia Novoradovskaya, Michael S Orr, Terry W Osborn, Adam Papallo, Tucker a Patterson, Roger G Perkins, Elizabeth H Peters, Ron Peterson, Kenneth L Philips, P Scott Pine, Lajos Pusztai, Feng Qian, Hongzu Ren, Mitch Rosen, Barry a Rosenzweig, Raymond R Samaha, Mark Schena, Gary P Schroth, Svetlana Shchegrova, Dave D Smith, Frank Staedtler, Zhenqiang Su, Hongmei Sun, Zoltan Szallasi, Zivana Tezak, Danielle Thierry-Mieg, Karol L Thompson, Irina Tikhonova, Yaron Turpaz, Beena Vallanat, Christophe Van, Stephen J Walker, Sue Jane Wang, Yonghong Wang, Russ Wolfinger, Alex Wong, Jie Wu, Chunlin Xiao, Qian Xie, Jun Xu, Wen Yang, Liang Zhang, Sheng Zhong, Yaping Zong, and William Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9):1151–61, September 2006.

[57] Jonathon Shlens. A Tutorial on Principal Component Analysis. *New York*, 2009.

[58] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):Article3, January 2004.

[59] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):Article3, January 2004.

[60] Gordon K Smyth, Matthew Ritchie, Natalie Thorne, and James Wettenhall. limma: Linear Models for Microarray Data. User's Guide.

[61] R. O. Duda P. E. Hart D. G. Stork. *Pattern classification.* John Wiley and Sons Inc., 2001.

[62] Lubert Stryer. *Biochemistry (Fourth Edition).* W.H. Freeman and Company, 1995.

[63] Ingenuity systems. Ingenuity Knowledge Base -page. http://www.ingenuity.com/science/knowledge_base.html, August 2012.

[64] Ingenuity systems. Ingenuity Pathway Analysis webpage. https://ingenuity.my.salesforce.com/home/home.jsp, July 2012.

[65] Kayoko Takashima, Yumiko Mizukawa, Katsumi Morishita, Manabu Okuyama, Toshihiko Kasahara, Naoki Toritsuka, Toshikazu Miyagishima, Taku Nagao, and Tetsuro Urushidani. Effect of the difference in vehicles on gene expression in the rat liver–analysis of the control data in the Toxicogenomics Project Database. *Life sciences*, 78(24):2787–96, May 2006.

[66] The Gene Ontology Consortium. Gene Ontology : Tool for the Unification of Biology. *Nature Genetics*, 25(may):25–29, 2000.

[67] Takeki Uehara, Atsushi Ono, Toshiyuki Maruyama, Ikuo Kato, Hiroshi Yamada, Yasuo Ohno, and Tetsuro Urushidani. The Japanese toxicogenomics project: application of toxicogenomics. *Molecular nutrition & food research*, 54(2):218–27, February 2010.

[68] Tetsuro Urushidani. Prediction of Hepatotoxicity Based on the Toxicogenomics Database. *Hepatotoxicity: From Genomics to in vitro and in vivo Models*, 2007.

[69] Tetsuro Urushidani and Taku Nagao. Toxicogenomics : Japanese Initiative. *Handbook of Toxicogenomics: Strategies and Applications, WILEY-VCH*, pages 623–631, 2005.

[70] W N Venables and B D Ripley. *Statistics Complements to Modern Applied Statistics with S, Fourth edition, Springer.* 2002.

[71] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq : a Revolutionary Tool for Transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, 2009.

[72] J F Waring, R Ciurlionis, R Jolly, M Heindel, and R G Ulrich. Microarray Analysis of Hepatotoxins *in vitro* Reveals a Correlation Between Gene Expression Profiles and Mechanisms of Toxicity. *Toxicology letters*, 120(1-3):359–68, March 2001.

[73] B. Alberts D. Bray J. Lewis M. Raff K. Roberts J.D. Watson. *Molecular Biology of the Cell (Second Edition).* Garland Publishing Inc., 1989.

[74] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M Lin, Vivian Peng, John Ngai, and Terence P Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research*, 30(4):e15, February 2002.

A. APPENDIXES

A.1 PCA code

```
# Order according to the t-statistic:
rtt = rowttests(Dataset, factor(Dataset$state))
ordtt <- order(rtt$p.value)
esTT = Dataset[ordtt[1:50],]

# Perform PCA and plot the results:
pc = prcomp(t(exprs(esTT)))
plot(pc$x)
```

A.2 KNN code

```
# Standardize the gene expression values:
rowIQRs = function(eSet) {
numSamp = ncol(eSet)
lowQ = row(eSet, floor(0.25 *numSamp))
upQ = row(eSet, ceiling(0.75  *numSamp))
upQ - lowQ
}
standardize = function(x) (x-rowMedians(x))/rowIQRs(x)
exprs(Dataset) = standardize(exprs(Dataset))

# To compute the distances:
eucD = dist(t(exprs(Dataset)))
eucM = as.matrix(eucD)

# To choose only 50 features:
Traintt = rowtttest(Dataset, ''state'')
ordTT = order(abs(Traintt$statistic), decreasing = TRUE)
Trainset = Trainset[ordTT[1:50]]

# Cross-validation:
train = t(exprs(Trainset))
cl = as.factor(Trainset$state)
```

```
knn_results = knn.cv(train, cl)

# To compare the results to known classes:
knntable <- table(Trainset$state, knn_results)
```

A.3 Tables and Figures

Table A.1: Top biofunctions, canonical pathways, upstream regulators, tox lists and networks from Ingenuitys Core analysis of differentially expressed genes from those samples with **necrosis** as histopathological finding. In Top Bio Functions and Top Tox Functions, the number of genes or molecules are presented in the parenthesis. In Canonical Pathways and Top Tox Lists the ratio molecules in that pathway that meet the cut-off criteria and the total number of molecules that make up that pathway is given. In Top Networks a network score representing the likelihood that the molecules are found by change is shown.

Necrosis-group (66 differentially expressed genes)	
Top Bio Function, Molecular and Cellular Functions	**Top Canonical Pathways**
Cell Cycle (11)	Aryl Hydrocarbon Receptor Signaling (8/159)
Drug Metabolism (13)	Metabolism of Xenobiotics by Cytochrome P450 (6/196)
Molecular Transport (25)	Retinol Metabolism (4/64)
Small Molecule Biochemistry (30)	Xenobiotic Metabolism Signaling (7/295)
Lipid Metabolism (22)	Pyruvate Metabolism (4/139)
Top Upstream regulators	**Top Tox Lists**
POR	Nongenotoxic Hepatocarcinogenicity Biomarker Panel (5/22)
PPARA (Activated)	Xenobiotic Metabolism Signaling (11/347)
AHR	Aryl Hydrocarbon Receptor Signaling (8/156)
methylprednisolone	CAR/RXR Activation (4/29)
NR1I3 (Activated)	Glutathione Depletion - CYP Induc. and React. Metab. (3/12)
Top Tox Functions, Hepatotoxicity	
Glutathione Depletion In Liver (3)	
Hepatocellular Carcinoma (7)	
Liver Hyperplasia/Hyperproliferation (7)	
Liver Fibrosis (3)	
Liver Steatosis (3)	
Top Networks	
Cell Cycle, Drug Metabolism, Glutathione Depletion In Liver (39)	
Small Molecule Biochemistry, Amino Acid Metabolism, Energy Production (34)	
Organ Development, Renal and Urological System Devel. and Func., Endocrine Sys. Disorders (21)	
Amino Acid Metabolism, Small Molecule Biochemistry, Nutritional Disease (21)	
Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry (19)	

Table A.2: Top biofunctions, canonical pathways, upstream regulators, tox lists and networks from Ingenuitys Core analysis of differentially expressed genes from those samples with **cellular change** as histopathological finding. In Top Bio Functions and Top Tox Functions, the number of genes or molecules are presented in the parenthesis. In Canonical Pathways and Top Tox Lists the ratio molecules in that pathway that meet the cut-off criteria and the total number of molecules that make up that pathway is given. In Top Networks a network score representing the likelihood that the molecules are found by change is shown.

Cellular change -group (274 differentially expressed genes)	
Top Bio Function, Molecular and Cellular Functions	**Top Canonical Pathways**
Lipid Metabolism (88)	Xenobiotic Metabolism Signaling(21/295)
Molecular Transport (104)	LPS/IL-1 Mediated Inhibition of RXR Func. (18/236)
Small Molecule Biochemistry (110)	Aryl Hydrocarbon Receptor Signaling (15/159)
Carbohydrate Metabolism (59)	Nitrogen Metabolism (8/119)
Vitamin and Mineral Metabolism (26)	Glycerolipid Metabolism (12/156)
Top Upstream regulators	**Top Tox Lists**
PPARA	Xenobiotic Metabolism Signaling (28/347)
ACOX1 (Inhibited)	Acute Renal Failure Panel (Rat) (13/62)
POR	LPS/IL-1 Mediated Inhibition of RXR Func. (21/246)
dexamethasone	Aryl Hydrocarbon Receptor Signaling (17/156)
pirinixic acid (Activated)	Nongenotoxic Hepatocarcinog. Biomarker Panel (8/22)
Top Tox Functions, Hepatotoxicity	
Liver Steatosis (19)	
Liver Necrosis/Cell Death (4)	
Liver Cholestasis (9)	
Hepatocellular Carcinoma (23)	
Liver Hyperplasia/Hyperproliferation (24)	
Top Networks	
Ophthalmic Disease, Hematological Disease, Metabolic Disease (42)	
Behavior, Nervous System Development and Function, Endocrine System Dev. and Func. (40)	
Drug Metabolism, Glutathione Depletion In Liver, Endocrine System Dev. and Func.(37)	
Cell-To-Cell Signaling and Interaction, Tissue Development, Lipid Metabolism (33)	
Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry (30)	

Table A.3: Top biofunctions, canonical pathways, upstream regulators, tox lists and networks from Ingenuitys Core analysis of differentially expressed genes from those samples with **cellular infiltration** as histopathological finding. In Top Bio Functions and Top Tox Functions, the number of genes or molecules are presented in the parenthesis. In Canonical Pathways and Top Tox Lists the ratio molecules in that pathway that meet the cut-off criteria and the total number of molecules that make up that pathway is given. In Top Networks a network score representing the likelihood that the molecules are found by change is shown.

Cellular infiltration -group (109 differentially expressed genes)	
Top Bio Function, Molecular and Cellular Functions	**Top Canonical Pathways**
Lipid Metabolism (38)	Nitrogen Metabolism (6/119)
Small Molecule Biochemistry (46)	Xenobiotic Metabolism Signaling (11/295)
Molecular Transport (44)	LPS/IL-1 Mediated Inhibition of RXR Function (9/236)
Cellular Development (15)	Metabolism of Xenobiotics by Cytochrome P450 (6/196)
Carbohydrate Metabolism (22)	PXR/RXR Activation (5/88)
Top Upstream regulators	**Top Tox Lists**
progesterone	Xenobiotic Metabolism Signaling (14/347)
TO-901317	LPS/IL-1 Mediated Inhibition of RXR Function (10/246)
dexamethasone	Nongenotoxic Hepatocarcinogenicity Biomarker Panel (4/22)
TNF	CAR/RXR Activation (4/29)
POR	NRF2-mediated Oxidative Stress Response (9/231)
Top Tox Functions, Hepatotoxicity	
Glutathione Depletion In Liver (3)	
Liver Cholestasis (5)	
Liver Steatosis (8)	
Hepatocellular Carcinoma (10)	
Liver Hyperplasia/Hyperproliferation (10)	
Top Networks	
Lipid Metabolism, Small Molecule Biochemistry, Molecular Transport (46)	
Hematological System Development and Function, Tissue Morphology, Cancer (36)	
Amino Acid Metabolism, Small Molecule Biochemistry, Lipid Metabolism (28)	
Small Molecule Biochemistry, Antigen Presentation, Cellular Movement (26)	
Ophthalmic Disease, Respiratory Disease, Hereditary Disorder (26)	

Table A.4: Top biofunctions, canonical pathways, upstream regulators, tox lists and networks from Ingenuitys Core analysis of differentially expressed genes from those samples with **vacuolar degeneration** as histopathological finding. In Top Bio Functions and Top Tox Functions, the number of genes or molecules are presented in the parenthesis. In Canonical Pathways and Top Tox Lists the ratio molecules in that pathway that meet the cut-off criteria and the total number of molecules that make up that pathway is given. In Top Networks a network score representing the likelihood that the molecules are found by change is shown.

Vacuolar degeneration -group (880 differentially expressed genes)	
Top Bio Function, Molecular and Cellular Functions	**Top Canonical Pathways**
Lipid Metabolism (201)	Fatty Acid Metabolism (26/183)
Small Molecule Biochemistry (253)	LPS/IL-1 Mediated Inhibition of RXR Func. (34/236)
Molecular Transport (164)	Arachidonic Acid Metabolism (21/206)
Energy Production (52)	Glycerolipid Metabolism (19/156)
Vitamin and Mineral Metabolism (80)	C21-Steroid Hormone Metabolism (8/66)
Top Upstream regulators	**Top Tox Lists**
pirinixic acid (Activated)	Fatty Acid Metabolism (28/123)
PPARA (Activated)	LPS/IL-1 Mediated Inhibition of RXR Function (40/246)
methylprednisolone	Cytochrome P450 Panel-Subst.:Fatty Acid (Mouse) (7/15)
clofibrate (Activated)	Cytochrome P450 Panel-Subst.:Fatty Acid (Rat) (7/12)
fenofibrate (Activated)	Acute Renal Failure Panel (Rat) (13/62)
Top Tox Functions, Hepatotoxicity	
Liver Hyperplasia/Hyperproliferation (55)	
Hepatocellular Carcinoma (54)	
Liver Steatosis (32)	
Liver Cholestasis (13)	
Liver Damage (18)	
Top Networks	
Cell Morphology, Embryonic Development, Organ Development (51)	
Lipid Metabolism, Small Molecule Biochemistry, Nucleic Acid Metabolism (38)	
Endocrine System Development and Function, Lipid Metabolism, Small Molecule Biochemistry (35)	
Cell Cycle, Nervous System Development and Function, Cell Death (35)	
Cell Cycle, DNA Replication, Recombination, and Repair, Cellular Assembly and Organization (34)	

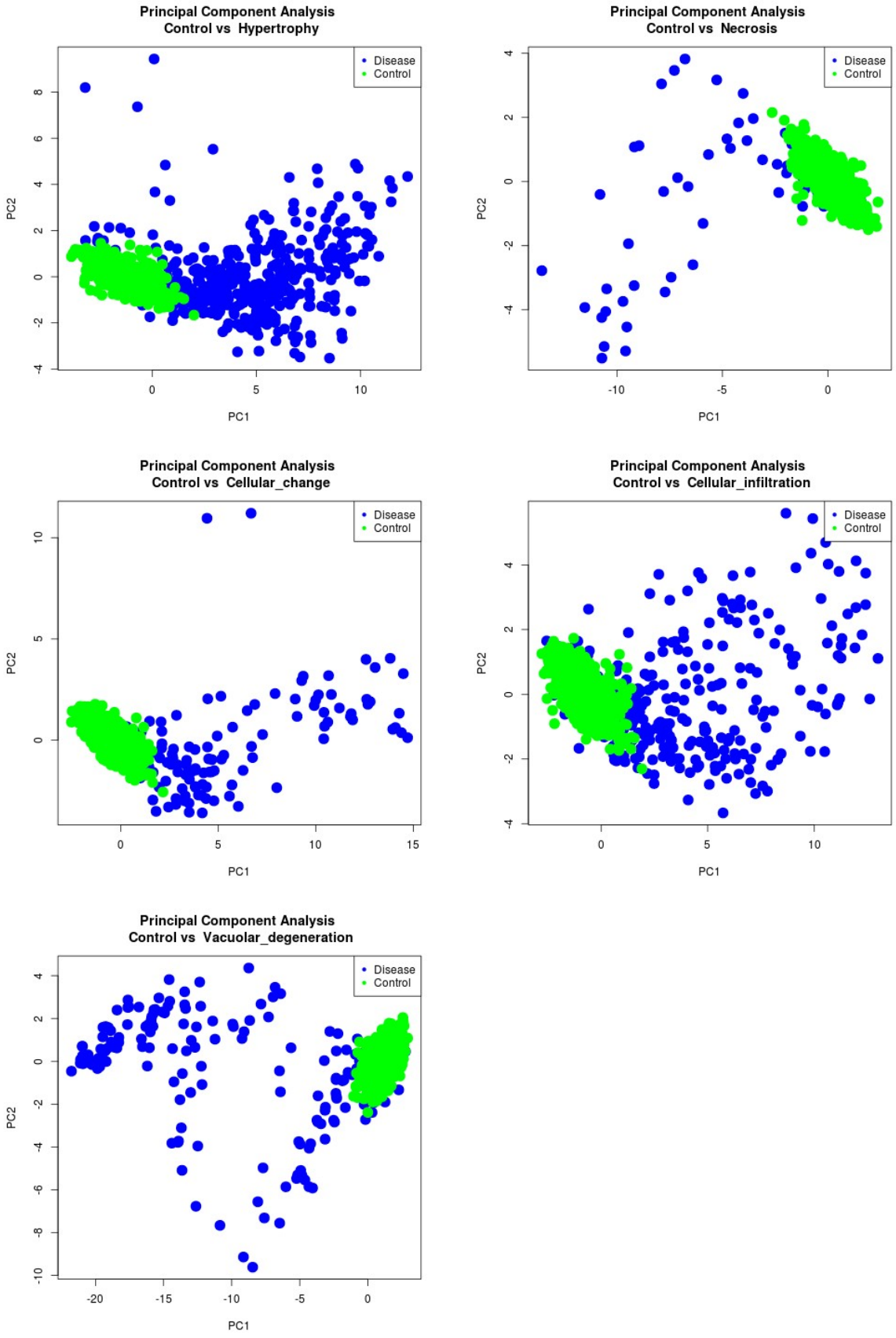


Figure A.1: PCA plots showing the separation between different histopathological groups (blue) and controls (green).

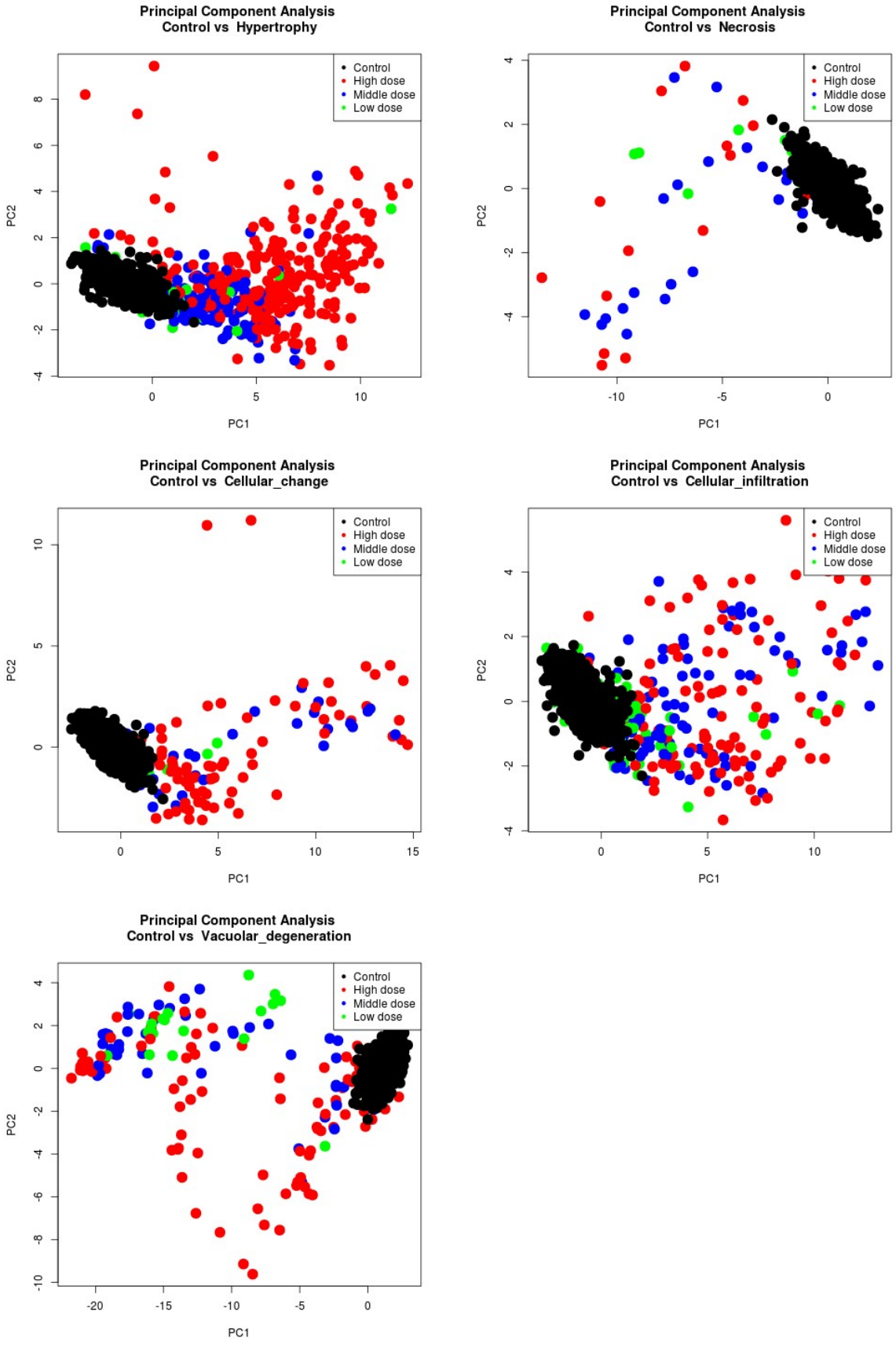


Figure A.2: PCA plots showing the separation between different histopathological groups colored according to the dose level and controls (controls). In each case, red indicated that the rat was treated with high dose, blue indicates middle dose and green low dose.

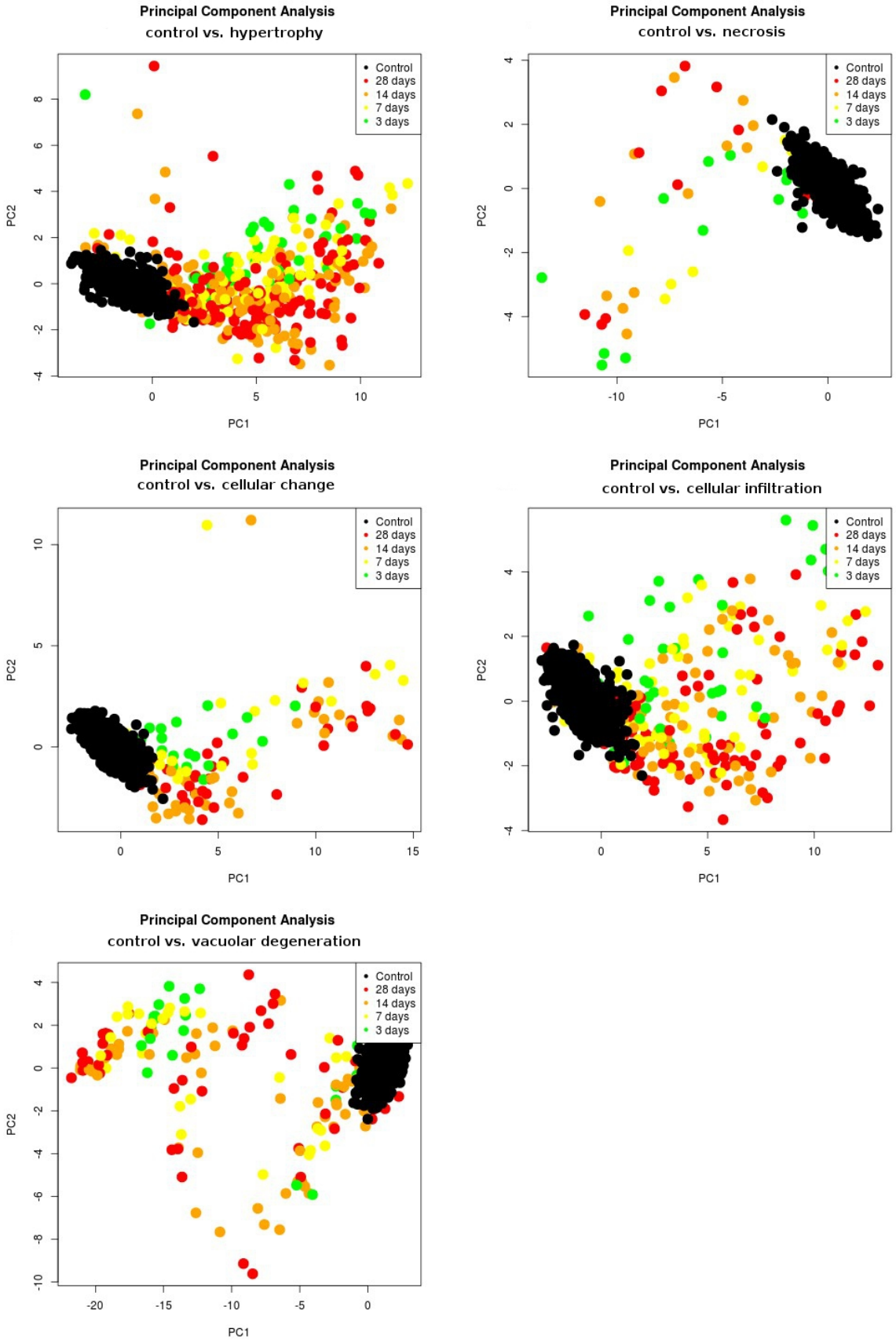


Figure A.3: PCA plots showing the separation between different histopathological groups colored according to the exposure time and controls (controls). In each case, red indicated that the rat was treated with a certain dose for 28 days, orange indicates exposure time of 14 days, yellow of 7 days and green three days.

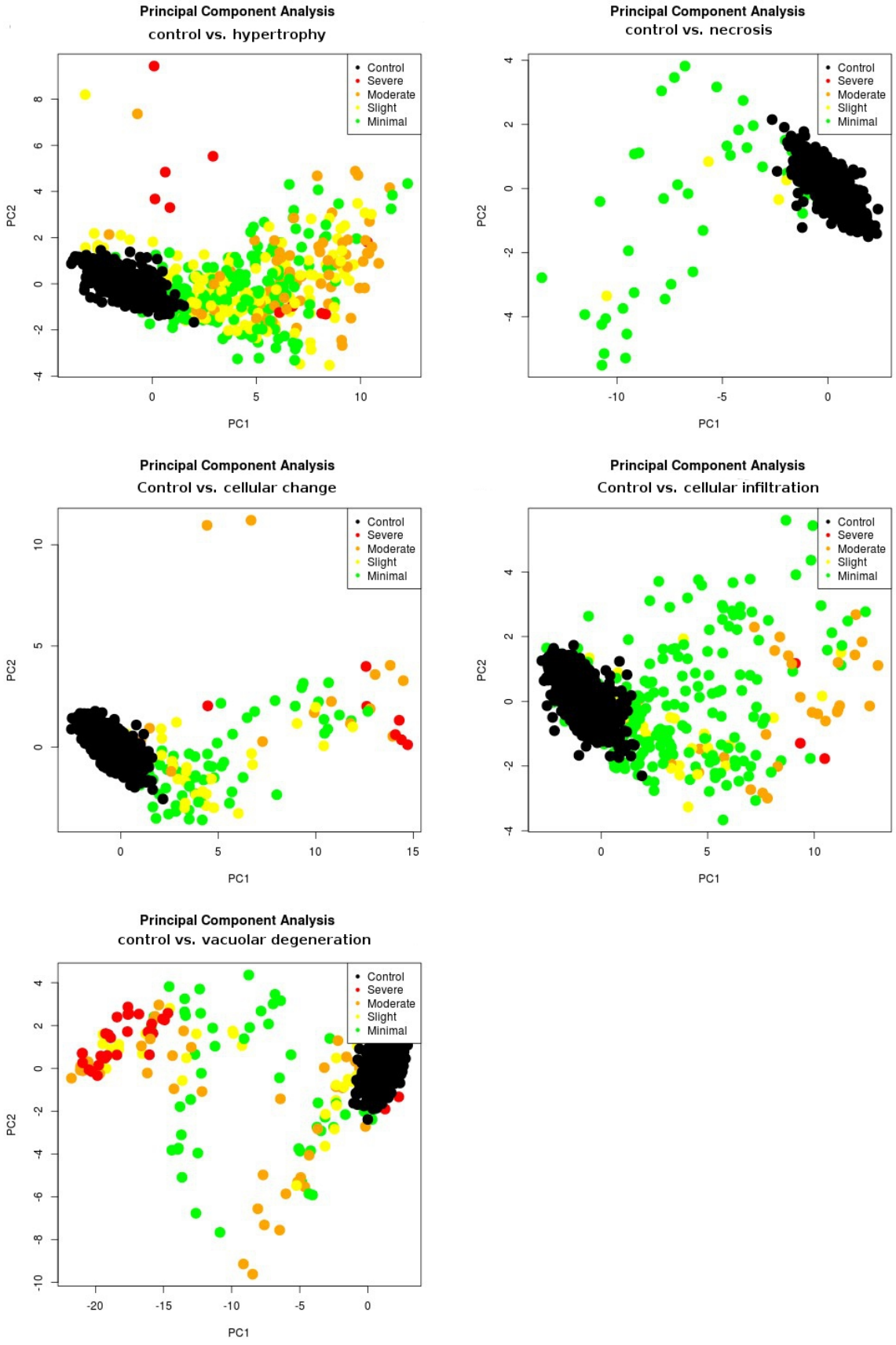


Figure A.4: PCA plots showing the separation between different histopathological groups colored according to the severity of the histopathological finding and controls (controls). In each case, red indicated that the finding was severe, orange indicates moderate finding, yellow slight and green minimal.

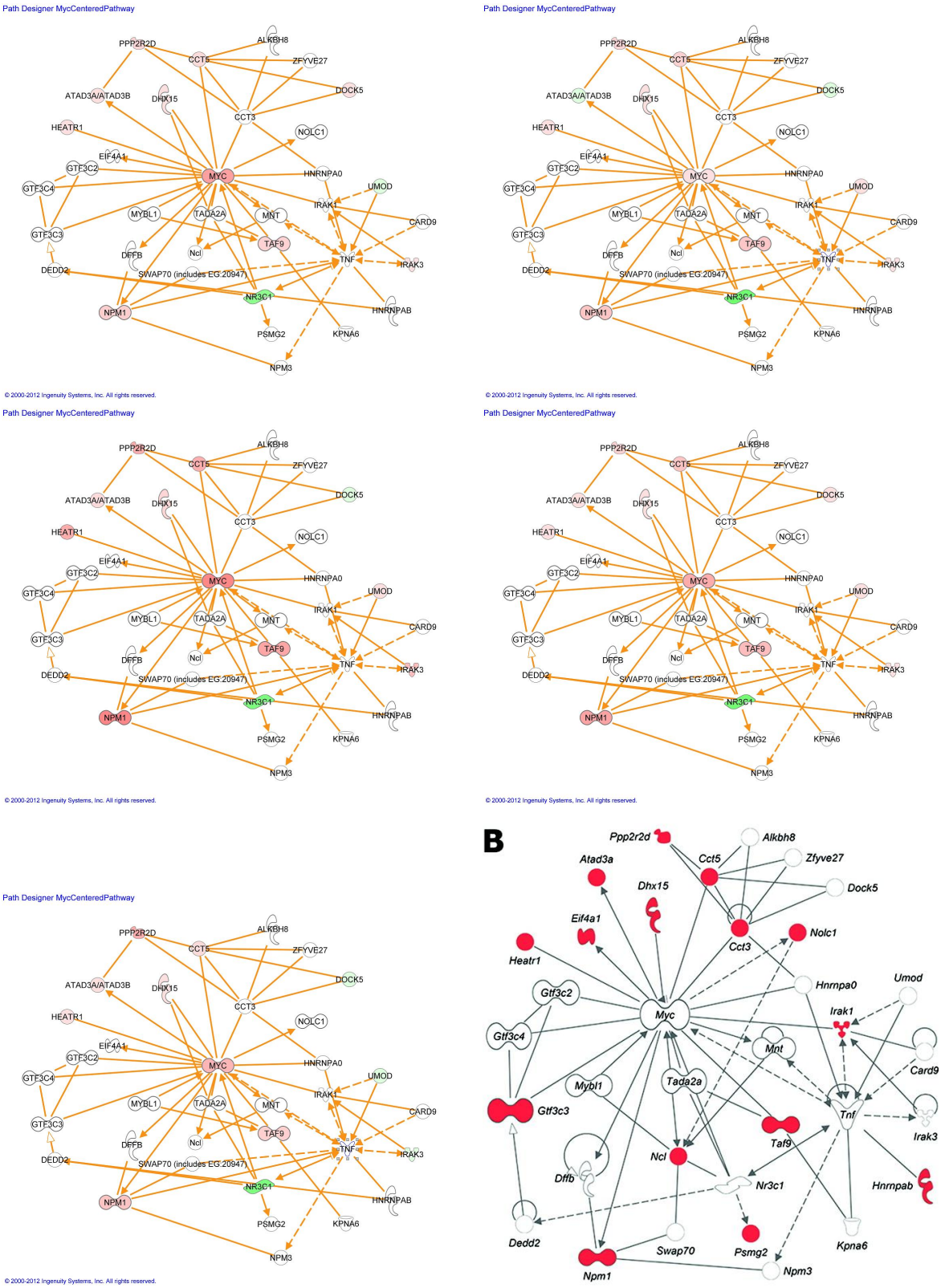


Figure A.5: Myc-centered molecular networks of different histopathological groups (form upper left corner: hypertrophy, necrosis, cellular change, cellular infiltration, vacuolar degeneration; images produced with Ingenuity) and of known hepatotoxic compounds (lowest right corner, image borrowed from [39]. Red ang green represent up- and down-regulated molecules, respectively.

Table A.5: Most interesting genes, their p-values and log fold-changes in **cellular change** -group.

Gene symbols	$\log_2$ FC	Adj. p-value	Gene names
Ces2c	1,081834	4,05E-160	carboxylesterase 2C
Stac3	-1,94251	4,43E-140	SH3 and cysteine rich domain 3
Bmf	-0,63551	5,29E-130	Bcl2 modifying factor
Aldh1a1	1,694144	2,55E-115	aldehyde dehydrogenase 1 family, member A1
Pvrl3	-0,32822	2,55E-115	poliovirus receptor-related 3
Gsta5	1,779908	7,10E-114	glutathione S-transferase Yc2 subunit
Pc	-0,37119	6,47E-111	pyruvate carboxylase
Slc46a1	-0,52542	4,66E-108	solute carrier family 46 (folate transporter), member 1
Nt5e	-0,40266	2,61E-104	5' nucleotidase, ecto
Apoa4	-0,50799	1,44E-102	apolipoprotein A-IV
Hsdl2	0,513135	1,35E-98	hydroxysteroid dehydrogenase like 2
Extl1	-0,50275	1,61E-98	exostoses (multiple)-like 1
Gulo	-0,41146	4,00E-94	gulonolactone (L-) oxidase
Gucy1b2	-0,50865	4,63E-94	guanylate cyclase 1, soluble, beta 2
S1pr1	-0,35374	1,63E-93	sphingosine-1-phosphate receptor 1
R3hdm2	-0,67071	1,35E-91	R3H domain containing 2
Mgat4b	-0,34883	1,67E-91	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme B
Slc25a13	-0,31765	3,07E-90	solute carrier family 25, member 13 (citrin)
Abcc6	-0,29062	1,01E-89	ATP-binding cassette, subfamily C (CFTR/MRP), member 6
Inhbe	-0,62306	8,32E-89	inhibin beta E
Aldh1l2	-0,4426	9,11E-89	aldehyde dehydrogenase 1 family, member L2
Inhbc	-0,505	3,41E-85	inhibin beta C
Nhp2	0,332523	3,95E-85	NHP2 ribonucleoprotein homolog (yeast)
Crem	-0,42877	1,07E-83	cAMP responsive element modulator
Npdc1	0,397115	1,67E-82	neural proliferation, differentiation and control, 1
Aig1	0,643931	1,85E-81	androgen-induced 1
Pklr	-0,62173	3,35E-81	pyruvate kinase, liver and RBC
Coq10a	-0,35931	3,48E-81	coenzyme Q10 homolog A (S. cerevisiae)
Faah	-0,31313	8,36E-81	fatty acid amide hydrolase
Aqp7	0,539105	1,52E-80	aquaporin 7
Rpp21	-0,29524	6,17E-80	ribonuclease P 21 subunit (human)
Cyp2t1	-0,4246	6,72E-80	cytochrome P450, family 2, subfamily t, polypeptide 1
Ugt1a1	0,277882	5,14E-79	UDP glucuronosyltransferase 1 family, polypeptide A1

Table A.6: Most interesting genes, their p-values and log fold-changes in **cellular infiltration** -group.

Gene symbols	$\log_2$ FC	Adj. p-value	Gene names
Aldh1a1	1,472385	4,60E-131	aldehyde dehydrogenase 1 family, member A1
Gsta5	1,445363	1,09E-125	glutathione S-transferase Yc2 subunit
Stac3	-1,27184	9,18E-112	SH3 and cysteine rich domain 3
R3hdm2	-0,54835	3,38E-107	R3H domain containing 2
Ces2c	0,839079	3,38E-107	carboxylesterase 2C
Ephx1	0,32881	2,84E-102	epoxide hydrolase 1, microsomal
App	0,40395	1,37E-98	amyloid beta (A4) precursor protein
Oat	-0,61323	4,36E-93	ornithine aminotransferase (gyrate atrophy)
Bmf	-0,36215	6,52E-93	Bcl2 modifying factor
Akr7a3	0,607537	2,27E-87	aldo-keto reductase family 7, member A3 (aflatoxin aldehyde reductase)
Cyp2t1	-0,28364	4,74E-79	cytochrome P450, family 2, subfamily t, polypeptide 1
Extl1	-0,31185	1,64E-77	exostoses (multiple)-like 1
Dexi	-0,29714	1,24E-75	dexamethasone-induced transcript
Cd276	0,277609	3,03E-75	Cd276 molecule
Apoa4	-0,40647	1,86E-73	apolipoprotein A-IV
Inhbe	-0,36692	2,17E-73	inhibin beta E
Mtmr7	-0,63384	2,75E-72	myotubularin related protein 7
Npdc1	0,29424	2,33E-71	neural proliferation, differentiation and control, 1
Pklr	-0,38811	6,33E-70	pyruvate kinase, liver and RBC
Pter	0,30234	3,64E-69	phosphotriesterase related
Gstm3	0,751076	1,14E-68	glutathione S-transferase mu 3
Prodh	-0,28227	3,96E-68	proline dehydrogenase
Sez6	-0,55664	3,61E-67	seizure related 6 homolog (mouse)
Hsdl2	0,298849	8,07E-67	hydroxysteroid dehydrogenase like 2
Zdhhc2	0,413422	1,35E-66	zinc finger, DHHC-type containing 2
Snx10	0,331591	2,58E-66	sorting nexin 10
LOC100363310	0,286244	1,44E-65	hypothetical protein LOC100363310
Lgals3	0,389899	2,25E-63	lectin, galactoside-binding, soluble, 3
Hal	-0,33319	3,13E-63	histidine ammonia lyase
Car3	-0,51823	3,79E-63	carbonic anhydrase 3
Dak	-0,29096	1,52E-62	dihydroxyacetone kinase 2 homolog (S. cerevisiae)
Nox4	-0,58841	1,49E-61	NADPH oxidase 4
Asns	0,545342	3,19E-61	asparagine synthetase
Vnn1	0,601442	7,52E-61	vanin 1
LOC367746	-0,69962	1,14E-60	similar to Spindlin-like protein 2 (SPIN-2)
Epcam	0,341203	2,93E-57	epithelial cell adhesion molecule
Ugt2b1	0,62223	3,50E-55	UDP glucuronosyltransferase 2 family, polypeptide B1
Gpx2	0,484635	1,90E-53	glutathione peroxidase 2

Table A.7: Most interesting genes, their p-values and log fold-changes in **hypertrophy** -group.

Gene symbols	$\log_2$ FC	Adj. p-value	Gene names
Gsta5	2,752421	0	glutathione S-transferase Yc2 subunit
Ces2c	1,782927	0	carboxylesterase 2C
Aldh1a1	2,509542	0	aldehyde dehydrogenase 1 family, member A1
Ugt1a1	0,505014	0,00E+00	UDP glucuronosyltransferase 1 family, polypeptide A1
Ephx1	0,617107	4,41E-276	epoxide hydrolase 1, microsomal
Pc	-0,45947	1,39E-257	pyruvate carboxylase
Hibch	0,454164	2,71E-247	3-hydroxyisobutyryl-Coenzyme A hydrolase
Mgst2	0,529633	1,75E-239	microsomal glutathione S-transferase 2
Gstm4	0,42989	3,80E-234	glutathione S-transferase mu 4
Gsr	0,51944	5,18E-214	glutathione reductase
App	0,541518	6,67E-207	amyloid beta (A4) precursor protein
Pvrl3	-0,34273	1,48E-206	poliovirus receptor-related 3
Pir	0,88573	9,77E-202	pirin (iron-binding nuclear protein)
Cltb	0,290364	9,13E-199	clathrin, light chain (Lcb)
Akr7a3	1,154102	1,37E-196	aldo-keto reductase family 7, member A3 (aflatoxin aldehyde reductase)
Dexi	-0,48441	7,63E-196	dexamethasone-induced transcript
Ugt2b1	1,379859	7,44E-181	UDP glucuronosyltransferase 2 family, polypeptide B1
Keg1	0,688626	2,12E-175	kidney expressed gene 1
Rexo2	0,276733	2,95E-166	REX2, RNA exonuclease 2 homolog (S. cerevisiae)
Hal	-0,49179	7,13E-164	histidine ammonia lyase
LOC498606	0,308457	1,25E-162	hypothetical protein LOC498606
Cyp3a23/3a1	0,403249	1,95E-162	cytochrome P450, family 3, subfamily a, polypeptide 23/polypeptide 1
H1f0	-0,38542	5,91E-161	H1 histone family, member 0
Csrp1	-0,49648	9,91E-159	cysteine and glycine-rich protein 1
Pter	0,454702	1,14E-157	phosphotriesterase related
Stac3	-1,34021	2,40E-154	SH3 and cysteine rich domain 3
Zdhhc2	0,636033	4,15E-154	zinc finger, DHHC-type containing 2
Nap1l1	0,275011	5,86E-153	nucleosome assembly protein 1-like 1
R3hdm2	-0,61979	2,63E-151	R3H domain containing 2
Cyp2t1	-0,39786	3,15E-150	cytochrome P450, family 2, subfamily t, polypeptide 1
Oat	-0,76416	9,10E-149	ornithine aminotransferase (gyrate atrophy)
Bmf	-0,39981	1,59E-147	Bcl2 modifying factor
LOC100363310	0,427139	2,83E-147	hypothetical protein LOC100363310
Gstm3	1,023933	8,94E-147	glutathione S-transferase mu 3
Cyp2j4	0,546376	4,12E-146	cytochrome P450, family 2, subfamily j, polypeptide 4
Car1	-0,47165	2,97E-141	carbonic anhydrase 1
Pklr	-0,59184	1,37E-140	pyruvate kinase, liver and RBC

Table A.8: Most interesting genes, their p-values and log fold-changes in **necrosis** -group.

Gene symbols	$\log_2$ FC	Adj. p-value	Gene names
Stac3	-1,11662	1,76E-73	SH3 and cysteine rich domain 3
Ces2c	0,746815	2,59E-69	carboxylesterase 2C
Gsta5	1,172307	4,48E-49	glutathione S-transferase Yc2 subunit
Ccng1	0,431957	3,69E-47	cyclin G1
Slc46a1	-0,3055	3,18E-44	solute carrier family 46 (folate transporter), member 1
Oat	-0,51692	3,90E-43	ornithine aminotransferase (gyrate atrophy)
Extl1	-0,32438	8,60E-43	exostoses (multiple)-like 1
Bmf	-0,30689	1,35E-42	Bcl2 modifying factor
Akr7a3	0,501073	1,97E-41	aldo-keto reductase family 7, member A3 (aflatoxin aldehyde reductase)
Ephx1	0,277169	4,38E-41	epoxide hydrolase 1, microsomal
R3hdm2	-0,45626	6,00E-41	R3H domain containing 2
Aldh1a1	0,999665	1,43E-38	aldehyde dehydrogenase 1 family, member A1
Cyp1a1	0,386277	4,91E-38	cytochrome P450, family 1, subfamily a, polypeptide 1
Apoa4	-0,32915	8,37E-38	apolipoprotein A-IV
Gstm3	0,643146	9,43E-38	glutathione S-transferase mu 3
Eci1	0,341017	2,67E-36	enoyl-Coenzyme A delta isomerase 1
Aqp7	0,344713	8,52E-36	aquaporin 7
Mdm2	0,320954	1,29E-34	Mdm2 p53 binding protein homolog (mouse)
Id1	0,338313	1,29E-34	inhibitor of DNA binding 1
Hsdl2	0,32662	5,14E-34	hydroxysteroid dehydrogenase like 2
Slc22a8	-0,45826	2,09E-33	solute carrier family 22 (organic anion transporter), member 8
Lgals3	0,378142	2,50E-33	lectin, galactoside-binding, soluble, 3
Csrp1	-0,30463	3,34E-33	cysteine and glycine-rich protein 1
Gdf15	0,36041	2,03E-32	growth differentiation factor 15
Pklr	-0,35306	2,67E-31	pyruvate kinase, liver and RBC
Sez6	-0,52018	5,76E-31	seizure related 6 homolog (mouse)
Car3	-0,4548	6,82E-30	carbonic anhydrase 3
Hapln3	0,314583	4,83E-28	hyaluronan and proteoglycan link protein 3
Aig1	0,329081	9,77E-28	androgen-induced 1
Inhbe	-0,30328	2,71E-27	inhibin beta E
Ech1	0,263608	9,00E-27	enoyl coenzyme A hydratase 1, peroxisomal
Rbp7	0,512751	6,41E-26	retinol binding protein 7, cellular
Olr59	-0,3263	8,49E-25	olfactory receptor 59
Cela1	-0,39367	1,10E-24	chymotrypsin-like elastase family, member 1
Serpina7	0,662639	1,14E-23	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 7
Nox4	-0,49581	8,64E-20	NADPH oxidase 4

Table A.9: Most interesting genes, their p-values and log fold-changes in **vacuolar degeneration** -group.

Gene symbols	$\log_2$ FC	Adj. p-value	Gene names
Ces2c	1,816776	0	carboxylesterase 2C
Ugt1a1	0,663839	0	UDP glucuronosyltransferase 1 family, polypeptide A1
Stac3	-3,23725	7,07E-308	SH3 and cysteine rich domain 3
Aldh1a1	2,791489	1,19E-295	aldehyde dehydrogenase 1 family, member A1
Acaa1a	1,478236	6,32E-266	acetyl-Coenzyme A acyltransferase 1A
Cyp2t1	-0,87097	7,42E-264	cytochrome P450, family 2, subfamily t, polypeptide 1
Pvrl3	-0,64039	6,40E-259	poliovirus receptor-related 3
Acot2	1,464696	3,64E-258	acyl-CoA thioesterase 2
Vnn1	2,920794	4,61E-256	vanin 1
Zdhhc2	1,115911	2,04E-252	zinc finger, DHHC-type containing 2
Hsdl2	1,520351	7,36E-251	hydroxysteroid dehydrogenase like 2
Npdc1	0,783117	1,23E-249	neural proliferation, differentiation and control, 1
Extl1	-0,88581	1,51E-247	exostoses (multiple)-like 1
Aig1	2,32903	1,71E-241	androgen-induced 1
Apoa4	-2,00383	4,41E-232	apolipoprotein A-IV
Csrp1	-0,92948	1,10E-231	cysteine and glycine-rich protein 1
Hibch	0,806046	1,96E-230	3-hydroxyisobutyryl-Coenzyme A hydrolase
Eci1	1,434926	8,06E-228	enoyl-Coenzyme A delta isomerase 1
Nap1l1	0,629533	2,56E-227	nucleosome assembly protein 1-like 1
Gulo	-0,59395	1,13E-226	gulonolactone (L-) oxidase
Gls2	-0,58743	2,78E-222	glutaminase 2 (liver, mitochondrial)
Hmgcl	0,672089	2,08E-217	3-hydroxymethyl-3-methylglutaryl-Coenzyme A lyase
Cyp2j4	1,147857	1,34E-216	cytochrome P450, family 2, subfamily j, polypeptide 4
Crat	1,635456	1,13E-215	carnitine acetyltransferase
Aqp7	2,050913	1,06E-214	aquaporin 7
Ehhadh	1,466583	2,11E-214	enoyl-Coenzyme A, hydratase/3-hydroxyacyl Coenzyme A dehydrogenase
Hal	-0,95475	8,15E-214	histidine ammonia lyase
Ech1	1,524938	9,89E-214	enoyl coenzyme A hydratase 1, peroxisomal
Mgat4b	-0,72416	3,03E-212	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme B
Me1	1,720769	1,46E-211	malic enzyme 1, NADP(+)-dependent, cytosolic
Cyp4a1	1,46288	1,46E-211	cytochrome P450, family 4, subfamily a, polypeptide 1

Table A.10: 50 first GO terms and corresponding p-values for cellular change -group.

GO term (BP)	p-value
cellular ketone metabolic process	5,38E-24
organic acid metabolic process	1,41E-23
carboxylic acid metabolic process	2,34E-23
oxoacid metabolic process	2,34E-23
lipid metabolic process	2,59E-21
monocarboxylic acid metabolic process	3,63E-20
response to organic substance	4,09E-19
response to hormone stimulus	8,09E-17
response to endogenous stimulus	4,12E-16
fatty acid metabolic process	5,68E-16
cellular lipid metabolic process	2,93E-15
response to external stimulus	3,11E-14
response to steroid hormone stimulus	6,59E-14
cellular response to chemical stimulus	7,71E-14
response to drug	2,92E-13
response to extracellular stimulus	3,66E-13
response to peptide hormone stimulus	1,97E-12
response to nutrient levels	4,98E-12
small molecule biosynthetic process	5,50E-12
cellular response to organic substance	2,53E-11
response to glucocorticoid stimulus	6,00E-11
oxidation-reduction process	7,79E-11
lipid biosynthetic process	1,32E-10
aging	1,61E-10
response to corticosteroid stimulus	1,94E-10
response to abiotic stimulus	5,45E-10
steroid metabolic process	1,28E-09
regulation of lipid biosynthetic process	1,70E-09
response to nutrient	2,57E-09
small molecule metabolic process	2,75E-09
response to organic cyclic compound	3,61E-09
regulation of lipid metabolic process	8,79E-09
response to stress	8,92E-09
cellular response to hormone stimulus	2,37E-08
organic acid biosynthetic process	2,44E-08
carboxylic acid biosynthetic process	2,44E-08
cellular amine metabolic process	2,76E-08
alcohol metabolic process	3,91E-08
response to toxin	6,20E-08
secondary metabolic process	7,83E-08
amine metabolic process	1,12E-07
response to inorganic substance	1,96E-07
fatty acid biosynthetic process	2,41E-07
cellular aromatic compound metabolic process	2,56E-07
cellular response to endogenous stimulus	2,86E-07
cellular response to peptide hormone stimulus	3,29E-07
very long-chain fatty acid metabolic process	5,09E-07
response to insulin stimulus	5,12E-07
metabolic process	7,42E-07

Table A.11: 50 first GO terms and corresponding p-values for cellular infiltration -group.

GO term (BP)	p-value
carboxylic acid metabolic process	6,81E-16
oxoacid metabolic process	6,81E-16
organic acid metabolic process	1,26E-15
cellular ketone metabolic process	1,83E-15
response to organic substance	1,24E-13
response to endogenous stimulus	5,35E-13
response to external stimulus	6,24E-12
monocarboxylic acid metabolic process	3,35E-11
response to hormone stimulus	2,16E-10
cellular response to chemical stimulus	2,96E-10
response to drug	6,63E-09
response to extracellular stimulus	7,46E-09
cellular response to organic substance	9,80E-09
response to nutrient levels	1,59E-08
response to stress	5,00E-08
response to nutrient	6,72E-08
fatty acid metabolic process	7,47E-08
cellular amino acid metabolic process	1,23E-07
lipid metabolic process	3,05E-07
cell death	5,52E-07
death	6,07E-07
cellular amine metabolic process	6,68E-07
response to chemical stimulus	7,75E-07
amine metabolic process	1,46E-06
apoptosis	1,56E-06
cellular response to hormone stimulus	1,60E-06
programmed cell death	1,99E-06
response to steroid hormone stimulus	2,15E-06
regulation of cell death	2,31E-06
negative regulation of catalytic activity	2,36E-06
cellular response to endogenous stimulus	2,95E-06
regulation of apoptosis	3,15E-06
regulation of programmed cell death	3,83E-06
response to vitamin	4,16E-06
small molecule biosynthetic process	4,24E-06
response to organic cyclic compound	5,75E-06
response to peptide hormone stimulus	5,97E-06
catabolic process	6,08E-06
regulation of catalytic activity	8,02E-06
phenylpropanoid metabolic process	1,06E-05
organic acid biosynthetic process	1,20E-05
carboxylic acid biosynthetic process	1,20E-05
regulation of response to external stimulus	1,36E-05
response to other organism	1,54E-05
secondary metabolic process	1,54E-05
negative regulation of cell death	2,37E-05
negative regulation of biological process	2,38E-05
regulation of biological quality	2,38E-05
aging	2,39E-05

Table A.12: 50 first GO terms and corresponding p-values for hypertrophy -group.

GO term (BP)	p-value
carboxylic acid metabolic process	1,13E-31
oxoacid metabolic process	1,13E-31
organic acid metabolic process	3,41E-31
cellular ketone metabolic process	6,64E-31
monocarboxylic acid metabolic process	2,81E-24
fatty acid metabolic process	1,79E-18
lipid metabolic process	4,01E-13
cellular lipid metabolic process	2,07E-11
organic acid biosynthetic process	9,26E-11
carboxylic acid biosynthetic process	9,26E-11
cellular amino acid metabolic process	1,38E-10
response to organic substance	2,79E-10
response to extracellular stimulus	6,31E-10
response to endogenous stimulus	7,74E-10
response to nutrient levels	1,01E-09
cellular amine metabolic process	2,94E-09
response to hormone stimulus	4,03E-09
small molecule biosynthetic process	1,04E-08
amine metabolic process	1,54E-08
small molecule metabolic process	2,29E-08
oxidation-reduction process	2,88E-08
response to external stimulus	3,16E-08
metabolic process	3,36E-08
response to xenobiotic stimulus	3,37E-08
response to steroid hormone stimulus	1,18E-07
response to nutrient	1,42E-07
xenobiotic metabolic process	2,30E-07
cellular response to xenobiotic stimulus	2,30E-07
cellular response to endogenous stimulus	3,16E-07
cellular response to chemical stimulus	4,53E-07
cellular response to hormone stimulus	4,66E-07
fatty acid biosynthetic process	5,33E-07
response to drug	1,03E-06
cellular metabolic process	1,78E-06
cellular response to organic substance	2,97E-06
lipid biosynthetic process	4,19E-06
xenobiotic catabolic process	7,22E-06
glutathione metabolic process	7,89E-06
organic acid catabolic process	8,39E-06
carboxylic acid catabolic process	8,39E-06
response to stress	9,16E-06
secondary metabolic process	1,14E-05
long-chain fatty acid metabolic process	1,38E-05
very long-chain fatty acid metabolic process	1,51E-05
peptide metabolic process	1,90E-05
cellular modified amino acid metabolic process	1,97E-05
cellular response to peptide hormone stimulus	2,20E-05
response to glucocorticoid stimulus	2,25E-05
pyruvate biosynthetic process	2,62E-05

Table A.13: 50 first GO terms and corresponding p-values for necrosis -group.

GO term (BP)	p-value
organic acid metabolic process	2,94E-14
carboxylic acid metabolic process	1,30E-13
oxoacid metabolic process	1,30E-13
cellular ketone metabolic process	2,93E-13
monocarboxylic acid metabolic process	4,01E-13
lipid metabolic process	5,04E-10
fatty acid metabolic process	3,60E-09
response to organic substance	2,62E-08
response to stress	4,81E-07
response to steroid hormone stimulus	4,95E-07
response to toxin	6,23E-07
response to external stimulus	9,18E-07
response to endogenous stimulus	1,19E-06
response to organic cyclic compound	2,45E-06
aging	2,88E-06
response to drug	4,44E-06
response to extracellular stimulus	4,81E-06
cellular lipid metabolic process	6,59E-06
response to estrogen stimulus	1,05E-05
response to other organism	1,97E-05
response to hormone stimulus	2,69E-05
oxidation-reduction process	3,28E-05
response to nutrient	3,47E-05
cellular response to chemical stimulus	3,67E-05
response to biotic stimulus	4,43E-05
secondary metabolic process	4,48E-05
response to nutrient levels	5,44E-05
toxin metabolic process	6,65E-05
response to abiotic stimulus	6,87E-05
response to bacterium	9,61E-05
multi-organism process	9,85E-05
response to vitamin	0,000176
vitamin metabolic process	0,000227
response to glucocorticoid stimulus	0,00027
cellular response to organic substance	0,000292
response to wounding	0,000306
response to vitamin A	0,000308
retinoic acid metabolic process	0,000361
response to corticosteroid stimulus	0,000387
response to iron(III) ion	0,000416
cellular response to radiation	0,000495
response to inorganic substance	0,000555
vitamin biosynthetic process	0,000594
response to chemical stimulus	0,000618
cellular hormone metabolic process	0,00066
endothelial cell chemotaxis	0,000689
response to oxidative stress	0,000836
regulation of anti-apoptosis	0,000871
response to ethanol	0,000959

Table A.14: 50 first GO terms and corresponding p-values for vacuolar degeneration -group.

GO term (BP)	p-value
cellular ketone metabolic process	2,66E-34
organic acid metabolic process	1,62E-33
carboxylic acid metabolic process	4,52E-33
oxoacid metabolic process	4,52E-33
monocarboxylic acid metabolic process	1,76E-26
fatty acid metabolic process	5,08E-22
lipid metabolic process	1,03E-20
cellular lipid metabolic process	1,41E-17
response to organic substance	6,19E-17
oxidation-reduction process	2,05E-16
metabolic process	4,00E-15
response to external stimulus	1,98E-14
response to endogenous stimulus	1,62E-13
small molecule metabolic process	3,24E-13
response to stress	8,86E-13
response to extracellular stimulus	1,72E-12
response to hormone stimulus	2,92E-12
catabolic process	6,52E-12
response to nutrient levels	1,42E-11
cellular metabolic process	6,22E-11
cellular catabolic process	7,85E-11
cellular response to chemical stimulus	1,44E-10
cellular amine metabolic process	1,59E-10
very long-chain fatty acid metabolic process	6,13E-10
cellular lipid catabolic process	7,75E-10
amine metabolic process	1,23E-09
response to nutrient	1,79E-09
organic acid biosynthetic process	1,80E-09
carboxylic acid biosynthetic process	1,80E-09
small molecule biosynthetic process	3,55E-09
cellular amino acid metabolic process	3,65E-09
lipid catabolic process	3,85E-09
alcohol metabolic process	4,43E-09
arachidonic acid metabolic process	5,77E-09
protein activation cascade	6,64E-09
organic acid catabolic process	7,09E-09
carboxylic acid catabolic process	7,09E-09
primary metabolic process	7,24E-09
organic ether metabolic process	8,30E-09
organic substance transport	1,10E-08
lipid biosynthetic process	1,26E-08
regulation of lipid metabolic process	1,28E-08
response to drug	1,61E-08
cellular response to organic substance	2,30E-08
response to steroid hormone stimulus	2,47E-08
fatty acid oxidation	2,52E-08
complement activation	3,02E-08
lipid oxidation	4,02E-08
small molecule catabolic process	8,55E-08

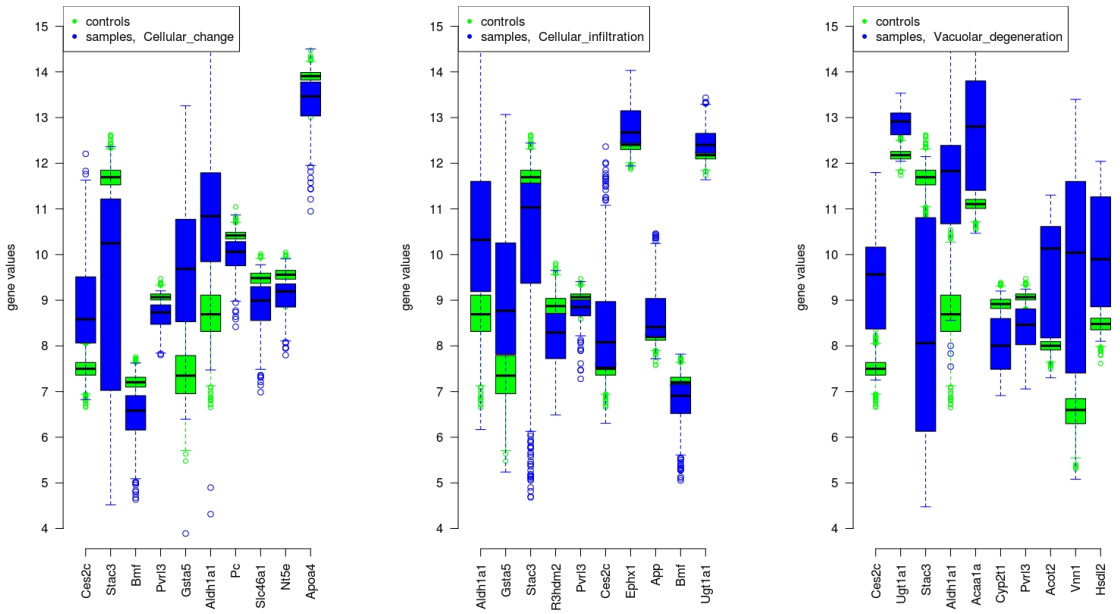


Figure A.6: Boxplots showing deviation of gene expression values of chosen genes in controls (green) and samples (blue) belonging to cellular change (left), cellular infiltration (middle) and vacuolar degeneration (right) groups. (The bottom and top of the box are the 25th and 75th percentile, the band near the middle of the box is the 50th percentile, and the whiskers show the most extreme data points which are no more than 1.5 times the interquartile range from the box. Data points beyond this range (outliers) are marked with circles.